On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms

Yan Bai*, Gareth O. Roberts†, and Jeffrey S. Rosenthal‡

(Last revised: July, 2010)

Abstract

This paper considers ergodicity properties of certain adaptive Markov chain Monte Carlo (MCMC) algorithms for multidimensional target distributions. It was previously shown in [23] that Diminishing Adaptation and Containment imply ergodicity of adaptive MCMC. We derive various sufficient conditions to ensure Containment.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms are widely used for approximately sampling from complicated probability distributions. However, it is often necessary to tune the scaling and other parameters before the algorithm will converge efficiently, and this can be very challenging especially in high dimensions. *Adaptive* MCMC algorithms attempt to overcome this challenge by learning from the past and modifying their transitions on the fly, in an effort to automatically tune the parameters and improve convergence. This approach was pioneered by the original *adaptive Metropolis* algorithm of Haario et al. [14], which can be viewed as a version of the Robbins-Monro stochastic control algorithm [20, 3]. Their paper was quickly followed by numerous other papers which generalised, modified, clarified, and proved theorems about various adaptive MCMC algorithms in various contexts and under various assumptions [7, 2, 15, 23, 24, 6, 1, 30, 32, 33, 11, 10, 4, 5, 8], as well as some general-purpose adaptive MCMC software [29, 31].

Despite this considerable progress, it remains true that verifying ergodicity of adaptive MCMC algorithms on unbounded state spaces remains non-trivial. Most of the ergodicity theorems assume a *Diminishing Adaptation* condition, whereby the amount of adapting done at iteration n converges to zero as $n \to \infty$, which is easily ensured by designing the algorithm appropriately. On a compact state space, this condition together with a simple continuity assumption suffices to ensure ergodicity of the algorithm (see e.g. Theorem 5 of [23]). However, on an unbounded state space, some additional assumption (such as the *Containment* condition discussed below) is also required or ergodicity may fail.

*Department of Statistics, University of Toronto, Toronto, ON M5S 3G3, CA. yanbai@utstat.toronto.edu

†Department of Statistics, University of Warwick, Coventry CV4 7AL, UK. gareth.o.roberts@warwick.ac.uk

‡Department of Statistics, University of Toronto, Toronto, ON M5S 3G3, CA. jeff@math.toronto.edu Supported in part by NSERC of Canada.

In this paper, we consider the Containment condition in more detail. In particular, we prove a number of results about sufficient (and occasionally necessary) conditions for Containment to hold. We hope that these results will allow users to verify Containment for adaptive algorithms more easily, and thus use adaptive MCMC more widely without fear of ergodicity problems.

1.1 Preliminaries

Consider a target distribution $\pi(\cdot)$ defined on a state space $\mathcal{X}$ with respect to some σ-field $\mathcal{B}(\mathcal{X})$ ($\pi(x)$ is also used to denote the density function). Let $\{P_\gamma : \gamma \in \mathcal{Y}\}$ be a family of transition kernels of time homogeneous Markov chains, each having the same stationary probability distribution π, i.e. $\pi P_\gamma = \pi$ for all $\gamma \in \mathcal{Y}$.

An *adaptive MCMC* algorithm $\mathbf{Z} := \{(X_n, \Gamma_n) : n \geq 0\}$ can be regarded as lying in the sample path space $\Omega := (\mathcal{X} \times \mathcal{Y})^\infty$. It proceeds as follows. We begin with an initial state $X_0 := x_0 \in \mathcal{X}$ and a kernel P_{Γ_0} where $\Gamma_0 := \gamma_0 \in \mathcal{Y}$. At each iteration $n+1$, X_{n+1} is generated from $P_{\Gamma_n}(X_n, \cdot)$, so that if $\mathcal{G}_n = \sigma(X_0, X_1, \ldots, X_n, \Gamma_0, \Gamma_1, \ldots, \Gamma_n)$, then for all $A \in \mathcal{B}(\mathcal{X})$,

$$\mathbb{P}_{(x_0,\gamma_0)}(X_{n+1} \in A \mid \mathcal{G}_n) = \mathbb{P}_{(x_0,\gamma_0)}(X_{n+1} \in A \mid X_n, \Gamma_n) = P_{\Gamma_n}(X_n, A)\,, \tag{1}$$

where $\mathbb{P}_{(x_0,\gamma_0)}$ represents the probabilities induced by our adaptive scheme when starting at $X_0 = x_0$ and $\Gamma_0 = \gamma_0$. Concurrently, Γ_{n+1} is obtained from some function of $X_0, \cdots, X_{n+1}$ and $\Gamma_0, \cdots, \Gamma_n$, according to the specific adaption scheme being used. (Intuitively, the adaptive scheme is designed so that it hopefully learns as it goes, so that the values Γ_n hopefully get correspondingly better, in terms of improved convergence of P_{Γ_n}, as n increases.)

In the paper, we study adaptive MCMC with the property Eq. (1). We say that the adaptive MCMC $\mathbf{Z}$ is *ergodic* if for any initial state $x_0 \in \mathcal{X}$ and any kernel index $\gamma_0 \in \mathcal{Y}$,

$$\lim_{n\to\infty} \left\| \mathbb{P}_{(x_0,\gamma_0)}(X_n \in \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} \;=\; 0\,,$$

where $\|\mu\|_{\mathrm{TV}} = \sup_{A\in\mathcal{B}(\mathcal{X})} |\mu(A)|$ is the usual total-variation metric on measures.

To study this ergodicity, we consider the properties of *Diminishing Adaptation* and *Containment*, following [23]. (There are several other closely related approaches to ergodicity of adaptive MCMC, see e.g. [2, 6, 30, 8].)

Diminishing Adaptation is the property that for any $X_0 = x_0$ and $\Gamma_0 = \gamma_0$, $\lim_{n\to\infty} D_n = 0$ in probability $\mathbb{P}_{(x_0,\gamma_0)}$ where $D_n = \sup_{x\in\mathcal{X}} \left\| P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot) \right\|_{\mathrm{TV}}$ represents the amount of adaptation performed between iterations n and $n+1$.

Containment is the property that for any $X_0 = x_0$ and $\Gamma_0 = \gamma_0$, for any $\epsilon > 0$, the stochastic process $\{M_\epsilon(X_n, \Gamma_n) : n \geq 0\}$ is bounded in probability $\mathbb{P}_{(x_0,\gamma_0)}$, i.e. for all $\delta > 0$, there is $N \in \mathbb{N}$ such that $\mathbb{P}_{(x_0,\gamma_0)}(M_\epsilon(X_n, \Gamma_n) \leq N) \geq 1 - \delta$ for all $n \in \mathbb{N}$, where $M_\epsilon(x,\gamma) = \inf\{n \geq 1 : \left\| P^n_\gamma(x,\cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} \leq \epsilon\}$ is the "ϵ-convergence time".

Theorem 1 ([23])**.** *Ergodicity of an adaptive MCMC algorithm is implied by Diminishing Adaptation and Containment.*

Thus, to ensure ergodicity of adaptive MCMC, it suffices to have Diminishing Adaptation and Containment. When designing adaptive algorithms, it is usually not difficult to ensure directly that Diminishing Adaptation holds. However, Containment may be more challenging, and is the subject of this paper.

Remark 1. *Atchadé et al. [8] allow for more general adaptive schemes, in which the different P_γ can have different stationary distributions, but we do not pursue that here.*

1.2 Organisation of the Paper

Section 2 below presents several examples to show that ergodicity can hold even if neither Containment nor Diminishing Adaptation holds, and that Diminishing Adaptation alone – even together with a weaker form of Containment – is not sufficient for ergodicity of adaptive MCMC. It also presents a simple *summable adaptive condition* which can be used to check ergodicity more easily. Finally, it discusses properties related to *simultaneous geometric ergodicity* which also imply ergodicity of adaptive algorithms.

Section 3 then discusses the weaker property of *simultaneous polynomial ergodicity*, and shows that this property also implies ergodicity of adaptive algorithms under appropriate conditions.

Section 4 specialises to adaptive algorithms based on families of Metropolis-Hastings algorithms. It shows that for lighter-than-exponential target distributions, ergodicity holds under relatively weak assumptions. On the other hand, for targets with exponential or hyperbolic tails, additional assumptions are required.

For ease of readability, all non-trivial proofs are deferred until Section 5.

2 Some Simple Results About Containment

We begin with a collection of relatively simple results about the Containment condition, before considering more substantial results in subsequent sections.

2.1 On Necessity of the Conditions

We begin with a very simple example to show that neither Diminishing Adaptation nor Containment are actually *necessary* for ergodicity of adaptive MCMC.

Example 1. *Let the state space $\mathcal{X} = \{1, 2\}$, and let the available Markov transition kernels be:*

$$P_\theta = \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}$$

for fixed $\theta \in (0,1)$. Obviously, for each $\theta \in (0,1)$, the stationary distribution is Unif($\mathcal{X}$), the uniform distribution on $\mathcal{X}$. Assume the following very simple state-independent adaptation scheme: at each time $n \geq 0$, we choose the transition kernel P_{θ_n}, where θ_n is some specific function of n.

Proposition 1. *For the adaptation scheme of Example 1, with $\theta_n = \frac{1}{(n+2)^r}$ for some fixed $r > 0$, we have the following:*
(i) *For any $r > 0$, Diminishing Adaptation holds but Containment does not;*
(ii) *If $r > 1$, then $\mu_0 P_{\theta_0} P_{\theta_1} \cdots P_{\theta_n} \to \mu$ where μ depends on μ_0, and in particular if $\mu_0 \neq \mathrm{Unif}(\mathcal{X})$ then $\mu \neq \mathrm{Unif}(\mathcal{X})$, i.e. the adaptive scheme is not ergodic.*
(iii) *If $0 < r \leq 1$, then for any probability measure μ_0 on $\mathcal{X}$, we have $\mu_0 P_{\theta_0} P_{\theta_1} \cdots P_{\theta_n} \to \mathrm{Unif}(\mathcal{X})$, i.e. the adaptive scheme is ergodic in this case.*

See the proof in Section 5.1.

Remark 2. *The chain in Proposition 1 is simply a time inhomogeneous Markov chain, artificially fit into the framework of adaptive MCMC. Although very simple, this example indicates the complexity of adaptive MCMC ergodicity. In particular:*
1. *For $r > 1$, the limiting distribution of the chain is not uniform. So it shows that Diminishing*

Adaptation alone cannot ensure ergodicity.
2. *For $0 < r \leq 1$, the algorithm is ergodic to the uniform distribution, but Containment does not hold. That is, although the "ϵ convergence time" goes to infinity (see Eq. (29)), the distance between the chain and the target is still decreasing to zero.*

Proposition 2. *For the adaptation scheme of Example 1, with $\theta_n = 1/2$ for n even, and $\theta_n = 1/n$ for n odd, both Diminishing Adaptation and Containment do not hold, but the chain still converges to the target distribution* $\mathrm{Unif}(\mathcal{X})$.

See the proof in Section 5.1.

Example 1 shows that Containment is not a strictly necessary condition for ergodicity to hold. In the following theorem, we prove that under certain additional conditions, Containment is in fact necessary for ergodicity of adaptive algorithms.

Theorem 2. *Suppose a family $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ has the property that there exists an increasing sequence of sets $\mathcal{D}_k \uparrow \mathcal{X}$ on the state space $\mathcal{X}$, such that for any $k > 0$,*

$$\lim_{n\to\infty} \sup_{\mathcal{D}_k\times\mathcal{Y}} \left\|P_\gamma^n(x,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}} = 0. \tag{2}$$

If an adaptive MCMC algorithm based on $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ is ergodic, then Containment holds.

Corollary 1. *Suppose that the parameter space $\mathcal{Y}$ is a metric space, and the adaptive scheme $\{\Gamma_n : n \geq 0\}$ is bounded in probability. Suppose that there exists an increasing sequence of sets $(\mathcal{D}_k, \mathcal{Y}_k) \uparrow \mathcal{X}\times\mathcal{Y}$ such that any $k > 0$,*

$$\lim_{n\to\infty} \sup_{\mathcal{D}_k\times\mathcal{Y}_k} \left\|P_\gamma^n(x,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}} = 0.$$

If the adaptive MCMC algorithm is ergodic then Containment holds.

For proofs of Theorem 2 and Corollary 1, see Section 5.2.

We now present a second, more complicated example. This example also fails to be ergodic, even though it satisfies Diminishing Adaptation, and also satisfies the "weak Containment" property that $\sup_{\gamma\in\mathcal{Y}}\sup_{x\in C} M_\epsilon(x,\gamma) < \infty$ for some small set C of positive stationary measure (indeed, that trivially holds for this example, with C any compact interval within $\mathcal{X}$, since $\mathcal{Y}$ is finite). Thus, this example shows that to ensure ergodicity, the full Containment condition is not redundant, and in particular it cannot simply be replaced by the "weak Containment" property.

Example 2. *Let the state space $\mathcal{X} = (0,\infty)$, and the kernel index set $\mathcal{Y} = \{-1, 1\}$. The target density $\pi(x) \propto \frac{\mathbb{I}(x>0)}{1+x^2}$ is a half-Cauchy distribution on the positive part of $\mathbb{R}$. At each time n, run the Metropolis-Hastings algorithm where the proposal value Y_n is generated by*

$$Y_n^{\Gamma_{n-1}} = X_{n-1}^{\Gamma_{n-1}} + Z_n \tag{3}$$

with i.i.d standard normal distribution $\{Z_n\}$, i.e. if $\Gamma_{n-1} = 1$ then $Y_n = X_{n-1} + Z_n$, while if $\Gamma_{n-1} = -1$ then $Y_n = \frac{1}{(1/X_{n-1})+Z_n}$. The adaptation is defined as

$$\Gamma_n = -\Gamma_{n-1}\mathbb{I}(X_n^{\Gamma_{n-1}} < \frac{1}{n}) + \Gamma_{n-1}\mathbb{I}(X_n^{\Gamma_{n-1}} \geq \frac{1}{n}), \tag{4}$$

i.e. we change Γ from 1 to -1 when $X < 1/n$, and change Γ from -1 to 1 when $X > n$, otherwise we do not change Γ.

Proposition 3. *The adaptive chain $\{X_n : n \geq 0\}$ defined in Example 2 is not ergodic, and Containment does not hold, although Diminishing Adaptation does hold.*

See the proof in Section 5.3.

2.2 Summable Adaptive Condition

In the following result, we use a simple coupling method to show that a certain summable adaptive condition implies ergodicity of adaptive MCMC.

Proposition 4. *Consider an adaptive MCMC $\{X_n : n \geq 0\}$ on the state space $\mathcal{X}$ with the kernel index space $\mathcal{Y}$. Under the following conditions:*
(i) *$\mathcal{Y}$ is finite. For any $\gamma \in \mathcal{Y}$, P_γ is ergodic with the stationary distribution π;*
(ii) *At each time n, Γ_n is a deterministic measurable function of $X_0, \cdots, X_n, \Gamma_0, \cdots, \Gamma_{n-1}$;*
(iii) *For any initial state $x_0 \in \mathcal{X}$ and any initial kernel index $\gamma_0 \in \mathcal{Y}$,*

$$\sum_{n=1}^{\infty} \mathbb{P}(\Gamma_n \neq \Gamma_{n-1} \mid X_0 = x_0, \Gamma_0 = \gamma_0) < \infty, \tag{5}$$

the adaptive MCMC $\{X_n : n \geq 0\}$ is ergodic with the stationary distribution π.

See the proof in Section 5.4.

Remark 3. *In Example 2, the transition kernel is changed when $X_n^{\Gamma_{n-1}}$ reaches below the bound $1/n$. If instead this bound is re-defined as $1/n^r$ for some $r > 1$, then Proposition 4 can be used (by adopting the procedure in Lemma 2 to check Eq. (5)) to show that the adaptive algorithm is ergodic.*

2.3 Simultaneous Geometric Drift Conditions Revisted

It was proven in [23] (see [2] for similar related results) that Containment is implied by *simultaneous strongly aperiodic geometric ergodicity* (S.S.A.G.E.). S.S.A.G.E. is the condition that there is $C \in \mathcal{B}(\mathcal{X})$, a function $V : \mathcal{X} \to [1, \infty)$, $\delta > 0, \lambda < 1$, and $b < \infty$, such that $\sup_{x \in C} V(x) < \infty$, and
(i) for each γ, $\exists$ a probability measure $\nu_\gamma(\cdot)$ on C with $P_\gamma(x, \cdot) \geq \delta \nu_\gamma(\cdot)$ for all $x \in C$, and
(ii) $P_\gamma V \leq \lambda V + b\mathbb{I}_C$ for all γ.

The idea of utilizing S.S.A.G.E. to check Containment is that S.S.A.G.E. guarantees there is a uniform quantitative bound of $\left\|P_\gamma^n(x, \cdot) - \pi(\cdot)\right\|_{\text{TV}}$ for all $\gamma \in \mathcal{Y}$. However, S.S.A.G.E. can in fact be weakened to the *simultaneously geometrically ergodic* condition (S.G.E.) studied by [27]. We say that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is S.G.E. if there is $C \in \mathcal{B}(\mathcal{X})$, some integer $m \geq 1$, a function $V : \mathcal{X} \to [1, \infty)$, $\delta > 0, \lambda < 1$, and $b < \infty$, such that $\sup_{x \in C} V(x) < \infty$, $\pi(V) < \infty$, and:
(i) C is a uniform ν_m-small set, i.e., for each γ, $\exists$ a probability measure $\nu_\gamma(\cdot)$ on C with $P_\gamma^m(x, \cdot) \geq \delta \nu_\gamma(\cdot)$ for all $x \in C$, and
(ii) $P_\gamma V \leq \lambda V + b\mathbb{I}_C$ for all γ.

Note that the difference between S.G.E. and S.S.A.G.E. is that a uniform minorization set C for all P_γ is assumed in S.S.A.G.E., however a uniform small set C is assumed in S.G.E. (see the definitions of minorization set and small set in [19, Chapter 5]).

Theorem 3. *S.G.E. implies Containment.*

See the proof in Section 5.5.

Corollary 2. *Consider the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ of Markov chains on $\mathcal{X} \subset \mathbb{R}^d$. Suppose that for any compact set $C \in \mathcal{B}(\mathcal{X})$, there exist some integer $m > 0$, $\delta > 0$ and a measure $\nu_\gamma(\cdot)$ on C for $\gamma \in \mathcal{Y}$ such that $P_\gamma^m(x, \cdot) \geq \delta \nu_\gamma(\cdot)$ for all $x \in C$. Suppose that there is a function $V : \mathcal{X} \to (1, \infty)$*

such that for any compact set $C \in \mathcal{B}(\mathcal{X})$, $\sup_{x \in C} V(x) < \infty$, $\pi(V) < \infty$, and

$$\limsup_{|x| \to \infty} \sup_{\gamma \in \mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < 1. \tag{6}$$

Then for any adaptive strategy using $\{P_\gamma : \gamma \in \mathcal{Y}\}$, Containment holds.

See the proof in Section 5.5.

3 Ergodicity via Simultaneous Polynomial Ergodicity

The previous section considered simultaneous *geometric* drift conditions. We now consider the extent to which Containment is ensured by the weaker property of simultaneous *polynomial* drift conditions.

3.1 Polynomial Ergodicity

There are many results available about polynomial ergodicity bounds for Markov chains [16, 17, 12, 13]. We begin by recalling in some detail a result by Fort and Moulines [13], giving a quantitative convergence bound for (non-adaptive) time-homogeneous Markov chains with polynomial (sub-geometric) convergence rates.

Theorem 4 ([13])**.** *Suppose that the time-homogeneous transition kernel P satisfies the following conditions:*

- *P is π-irreducible for an invariant probability measure π;*
- *There exist some sets $C \in \mathcal{B}(\mathcal{X})$ and $D \in \mathcal{B}(\mathcal{X})$, $C \subset D$, $\pi(C) > 0$ and an integer $m \geq 1$, such that for any $(x, x') \in \triangle := C \times D \cup D \times C$, $A \in \mathcal{B}(\mathcal{X})$,*

$$P^m(x, A) \wedge P^m(x', A) \geq \rho_{x,x'}(A) \tag{7}$$

  where $\rho_{x,x'}$ is some measure on $\mathcal{X}$ for $(x, x') \in \triangle$, and $\epsilon^- := \inf_{(x,x') \in \triangle} \rho_{x,x'}(\mathcal{X}) > 0$.
- *Let $q \geq 1$. There exist some measurable functions $V_k : \mathcal{X} \to \mathbb{R}^+ \setminus \{0\}$ for $k \in \{0, 1, \ldots, q\}$, and for $k \in \{0, 1, \ldots, q-1\}$, for some constants $0 < a_k < 1$, $b_k < \infty$, and $c_k > 0$ such that*

$$\begin{gathered} PV_{k+1}(x) \leq V_{k+1}(x) - V_k(x) + b_k \mathbb{I}_C(x), \inf_{x \in \mathcal{X}} V_k(x) \geq c_k > 0, \\ V_k(x) - b_k \geq a_k V_k(x), x \in D^c, \\ \sup_D V_q < \infty. \end{gathered} \tag{8}$$

- *$\pi(V_q^\beta) < \infty$ for some $\beta \in (0, 1]$.*

Then, for any $x \in \mathcal{X}$, $n \geq m$,

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \leq \min_{1 \leq l \leq q} B_l^{(\beta)}(x, n), \tag{9}$$

with

$$B_l^{(\beta)}(x,n) = \frac{\epsilon^+ \left\langle (I - A_m^{(\beta)})^{-1} \delta_x \otimes \pi(W^\beta), e_l \right\rangle}{S(l, n+1-m)^\beta + \sum_{j \geq n+1-m} (1-\epsilon^+)^{j-(n-m)} (S(l,j+1)^\beta - S(l,j)^\beta)},$$

where $< \cdot, \cdot >$ *denotes the inner product in* $\mathbb{R}^{q+1}$, $\{e_l\}$, $0 \leq l \leq q$ *is the canonical basis on* $\mathbb{R}^{q+1}$, I *is the identity matrix;*

$$\delta_x \otimes \pi(W^\beta) := \int \delta_x(dy)\pi(dy')W^\beta(y,y')$$

where $W^\beta(x,x') := \left(W_0^\beta(x,x'), \cdots, W_q^\beta(x,x')\right)^T$ *with* $W_0(x,x') := 1$ *and*

$$W_l(x,x') = \mathbb{I}_\triangle(x,x') + \mathbb{I}_{\triangle^c}(x,x') \left(\prod_{k=0}^{l-1} a_k\right)^{-1} (m(V_0))^{-1} (V_l(x) + V_l(x')) \text{ for } 1 \leq l \leq q$$

where $m(V_0) := \inf_{(x,x') \in \triangle^c} \{V_0(x) + V_0(x')\}$*;*

$$S(0,k) := 1 \text{ and } S(i,k) := \sum_{j=1}^{k} S(i-1,j), i \geq 1;$$

$$A_m^{(\beta)} := \begin{pmatrix} A_m^{(\beta)}(0) & 0 & \cdots & 0 & 0 \\ A_m^{(\beta)}(1) & A_m^{(\beta)}(0) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_m^{(\beta)}(q-1) & A_m^{(\beta)}(q-2) & \cdots & A_m^{(\beta)}(0) & 0 \\ A_m^{(\beta)}(q) & A_m^{(\beta)}(q-1) & \cdots & A_m^{(\beta)}(1) & A_m^{(\beta)}(0) \end{pmatrix},$$

where $A_m^{(\beta)}(l) := \sup_{(x,x') \in \triangle} \sum_{i=0}^{l} S(i,m)^\beta \left(1 - \rho_{x,x'}(\mathcal{X})\right) \int R_{x,x'}(x,dy) R_{x,x'}(x',dy') W_{l-i}^\beta(y,y')$, *where the residual kernel*

$$R_{x,x'}(u,dy) := \left(1 - \rho_{x,x'}(\mathcal{X})\right)^{-1} \left(P_\gamma^m(u,dy) - \rho_{x,x'}(dy)\right);$$

and $\epsilon^+ := \sup_{(x,x') \in \triangle} \rho_{x,x'}(\mathcal{X})$.

Remark 4. *In the* $B_l^{(\beta)}(x,n)$, ϵ^+ *depends on the set* $\triangle$ *and the measure* $\rho_{x,x'}$*; the matrix* $(I - A_m^{(\beta)})^{-1}$ *depends on the set* $\triangle$*, the transition kernel* P, $\rho_{x,x'}$ *and the test functions* V_k*;* $\delta_x \otimes \pi(W^\beta)$ *depends on the set* $\triangle$ *and the test functions* V_k*.*
Consider the special case of the theorem: $\rho_{x,x'}(dy) = \delta\nu(dy)$ *where* ν *is a probability measure with* $\nu(C) > 0$*, and* $\triangle := C \times C$*.*
1. $\epsilon^+ = \epsilon^- = \delta$*.*
2. $I - A_m^{(\beta)}$ *is a lower triangle matrix so* $(I - A_m^{(\beta)})^{-1} = \left(b_{ij}^{(\beta)}\right)_{i,j=1,\ldots,q+1}$ *is also a lower triangle matrix, and fixing* $k \geq 0$ *all* $b_{i,i-k}^{(\beta)}$ *are equal.* $b_{ii}^{(\beta)} = \frac{1}{1 - A_m^{(\beta)}(0)}$*. For* $i > j$, $b_{ij}^{(\beta)}$ *is the polynomial combination of* $A_m^{(\beta)}(0), \cdots, A_m^{(\beta)}(i+1)$ *divided by* $(1 - A_m^{(\beta)}(0))^i$*. By some algebra, we can obtain that* $b_{21}^{(\beta)} = \frac{A_m^{(\beta)}(1)}{(1 - A_m^{(\beta)}(0))^2}$*. So, by calculating* $B_1^{(\beta)}(x,n)$*, we can get the quantitative bound with a simple form.* $B_1^{(\beta)}(x,n)$ *only involves two test functions* $V_0(x)$ *and* $V_1(x)$*.*

Remark 5. *From Equation (8), $V_0(x) \geq b_0/(1-\alpha_0) > b_0$ because $0 < \alpha_0 < 1$. Consider the drift condition: $PV_1 - V_1 \leq -V_0 + b_0\mathbb{I}_C$. Since $\pi P = \pi$, $\pi(V_0) \leq b_0\pi(C) \leq b_0$. Hence, the V_0 in the theorem can not be constant.*

Remark 6. *Without the condition $\pi(V^\beta) < \infty$, the bound in Equation (9) can also be obtained. However, the bound is possibly infinity. The subscript l of $B_l^{(\beta)}(x,n)$ and β can explain the polynomial rate. The related rate is $S(l, n+1-m)^\beta = \mathrm{O}((n+1-m)^{l\beta})$. It can be observed that $B_l^{(\beta)}(x,n)$ involves test functions $V_0(x), \cdots, V_l(x)$, and $\limsup_n n^{\beta l} B_l^{(\beta)}(x,n) < \infty$. The maximal rate of convergence is equal to $q\beta$.*

3.2 Polynomial Ergodicity and Adaptive MCMC

To prove Containment using polynomial ergodicity, we shall require some additional assumptions, as follows. Say that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is *simultaneously polynomially ergodic* (S.P.E.) if the conditions (A1)-(A4) are satisfied.

A1: each P_γ is ϕ_γ-irreducible with stationary distribution $\pi(\cdot)$;

Remark 7. *By Proposition 10.1.2 of [19], if P_γ is φ-irreducible, then P_γ is π-irreducible and the invariant measure π is a maximal irreducibility measure.*

A2: there is a set $C \subset \mathcal{X}$, some integer $m \in \mathbb{N}$, some constant $\delta > 0$, and some probability measure $\nu_\gamma(\cdot)$ on $\mathcal{X}$ such that:

$$\pi(C) > 0, \text{ and } P_\gamma^m(x,\cdot) \geq \delta\mathbb{I}_C(x)\nu_\gamma(\cdot) \text{ for all } x \in \mathcal{X},\ \gamma \in \mathcal{Y}; \tag{10}$$

Remark 8. *In Theorem 4, there is one condition Eq. (7) ensuring the splitting technique. Here we consider the special case of that condition: $\rho_{x,x'}(dy) = \delta\nu_\gamma(dy)$ and $\triangle = C \times C$. Thus, by Remark 4, the bound of $\left\|P_\gamma^n(x,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}}$ depends on C, m, the minorization constant δ, $\pi(\cdot)$, ν_γ, and test functions $V_l(x)$ so we assume that they are uniform on all the transition kernels.*

A3: there is $q \in \mathbb{N}$ and measurable functions: $V_0, V_1, \ldots, V_q : \mathcal{X} \to (0,\infty)$ where $V_0 \leq V_1 \leq \cdots \leq V_q$, such that for $k = 0, 1, \ldots, q-1$, there are $0 < \alpha_k < 1$, $b_k < \infty$, and $c_k > 0$ such that:

$$P_\gamma V_{k+1}(x) \leq V_{k+1}(x) - V_k(x) + b_k\mathbb{I}_C(x),\ V_k(x) \geq c_k \text{ for } x \in \mathcal{X} \text{ and } \gamma \in \mathcal{Y}; \tag{11}$$

$$V_k(x) - b_k \geq \alpha_k V_k(x) \text{ for } x \in \mathcal{X}/C; \tag{12}$$

$$\sup_{x \in C} V_q(x) < \infty. \tag{13}$$

Remark 9. *For $x \in C$, $\nu_\gamma(V_l) \leq \frac{1}{\delta}P_\gamma^m V_l(x) \leq \frac{1}{\delta}\sup_{x\in C} V_l(x) + \frac{mb_{l-1}}{\delta}$.*

A4: $\pi(V_q^\beta) < \infty$ for some $\beta \in (0,1]$.

In terms of these assumptions, we have the following.

Theorem 5. *Suppose an adaptive MCMC algorithm satisfies Diminishing Adaptation. Then, the algorithm is ergodic under any of the following conditions:*
(i) *S.P.E., and the number q of simultaneous drift conditions is strictly greater than two;*

(ii) *S.P.E., and when the number q of simultaneous drift conditions is greater than or equal to two, there exists an increasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that $V_1(x) \le f(V_0(x))$;*
(iii) *Under the conditions* (A1) *and* (A2)*, there exist some positive constants $c > 0$, $b' > b > 0$, $\alpha \in (0,1)$, and a measurable function $V(x) : \mathcal{X} \to \mathbb{R}+$ with $V(x) \ge 1$ and $\sup_{x\in C} V(x) < \infty$ such that*

$$P_\gamma V(x) - V(x) \le -cV^\alpha(x) + b\mathbb{I}_C(x) \text{ for all } x \in \mathcal{X},\ \gamma \in \mathcal{Y}, \tag{14}$$

and $cV^\alpha(x)\mathbb{I}_{C^c}(x) \ge b'$ for all $x \in \mathcal{X}$;
(iv) *Under the condition* (A1), (A2), *and* (A4)*, there exist some constant $b' > b > 0$, two measurable functions $V_0 : \mathcal{X} \to \mathbb{R}^+$ and $V_1 : \mathcal{X} \to \mathbb{R}^+$ with $1 \le V_0(x) \le V_1(x)$ and $\sup_{x\in C} V_1(x) < \infty$ such that*

$$P_\gamma V_1(x) - V_1(x) \le -V_0(x) + b\mathbb{I}_C(x) \text{ for all } x \in \mathcal{X},\ \gamma \in \mathcal{Y}, \tag{15}$$

and $V_0(x)\mathbb{I}_{C^c}(x) \ge b'$ for all $x \in \mathcal{X}$, and the process $\{V_1(X_n) : n \ge 0\}$ is bounded in probability.

For a proof of Theorem 5, see Section 5.6.

Remark 10. *In part (iii), (A4) is then implied by Theorem 14.3.7 of [19], with $\beta = \alpha$.*

Remark 11. *Atchadé and Fort [6] recently proved a result closely related to the above, using a coupling method similar to that in [23]. Their Corollary 2.2 establishes ergodicity of adaptive MCMC algorithms under the assumptions of uniform strong aperiodicity, simultaneous drift conditions of the form (13), and uniform convergence on any sublevel set of the test function $V(x)$. Thus, they essentially reprove part (iii) of our Theorem 5, but under somewhat different assumptions.*

4 Ergodicity of Adaptive Metropolis-Hastings algorithms

We now present some ergodicity results for various adaptive Metropolis-Hastings algorithms. (For similar results about adaptive Metropolis-within-Gibbs algorithms, see [9].)

4.1 General Adaptive Metropolis-Hastings algorithms

We first consider general Metropolis-Hastings algorithms. We begin with some notation.

The target density $\pi(\cdot)$ is defined on the state space $\mathcal{X} \subset \mathbb{R}^d$. In what follows, we shall write $\langle \cdot, \cdot \rangle$ for the usual scalar product on $\mathbb{R}^d$, $|\cdot|$ for the Euclidean and the operator norm, $n(z) := z/|z|$ for the normed vector of z, ∇ for the usual differential (gradient) operator, $m(x) := \nabla\pi(x)/|\nabla\pi(x)|$, $\mathrm{B}^d(x, r) = \{y \in \mathbb{R}^d : |y - x| < r\}$ for the hyperball on $\mathbb{R}^d$ with the center x and the radius r, $\bar{\mathrm{B}}^d(x, r)$ for the closure of the hyperball, and $\mathrm{Vol}(A)$ for the volume of the set $A \subset \mathbb{R}^d$.

Say an adaptive MCMC is an *Adaptive Metropolis-Hastings algorithm* if each kernel P_γ is a Metropolis-Hastings algorithm, i.e. is of the form

$$P_\gamma(x, dy) = \alpha_\gamma(x, y)Q_\gamma(x, dy) + \left[1 - \int_{\mathcal{X}} \alpha_\gamma(x, z)Q_\gamma(x, dz)\right] \delta_x(dy) \tag{16}$$

where $Q_\gamma(x, dy)$ is the proposal distribution, $\alpha_\gamma(x, y) := \left(\frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)} \wedge 1\right) \mathbb{I}(y \in \mathcal{X})$, and μ_d is Lebesgue measure. Say an adaptive Metropolis-Hastings algorithm is an *Adaptive Metropolis algorithm* if each $q_\gamma(x, y)$ is symmetric, i.e. $q_\gamma(x, y) = q_\gamma(x - y) = q_\gamma(y - x)$.

[16] give conditions which imply geometric ergodicity of symmetric random-walk-based Metropolis algorithm on $\mathbb{R}^d$ for target distribution with lighter-than-exponential tails, [see other related results in 18, 25]. Here, we extend their result a little for target distributions with exponential tails.

Definition 1 (Lighter-than-exponential tail). *The density $\pi(\cdot)$ on $\mathbb{R}^d$ is lighter-than-exponentially tailed if it is positive and has continuous first derivatives such that*

$$\limsup_{|x|\mapsto\infty} \langle n(x), \nabla \log \pi(x) \rangle = -\infty. \tag{17}$$

Remark 12. 1. *The definition implies that for any $r > 0$, there exists $R > 0$ such that*

$$\frac{\pi(x + \alpha n(x)) - \pi(x)}{\pi(x)} \leq -\alpha r, \text{ for } |x| \geq R, \alpha > 0.$$

It means that $\pi(x)$ is exponentially decaying along any ray, but with the rate r tending to infinity as x goes to infinity.
2. *The normed gradient $m(x)$ will point towards the origin, while the direction $n(x)$ points away from the origin. For Definition 1, $\langle n(x), \nabla \log \pi(x) \rangle = \frac{|\nabla \pi(x)|}{\pi(x)} \langle n(x), m(x) \rangle$. Even $\limsup\limits_{|x|\mapsto\infty} \langle n(x), m(x) \rangle < 0$, Eq. (17) might not be true. E.g. $\pi(x) \propto \frac{1}{1+x^2}$, $x \in \mathbb{R}$. $m(x) = -n(x)$ so that $\langle n(x), m(x) \rangle = -1$. $\langle n(x), \nabla \log \pi(x) \rangle = -\frac{2|x|}{1+x^2}$ so $\lim\limits_{|x|\mapsto\infty} \langle n(x), \nabla \log \pi(x) \rangle = 0$.*

Definition 2 (Exponential tail). *The density function $\pi(\cdot)$ on $\mathbb{R}^d$ is exponentially tailed if it is a positive, continuously differentiable function on $\mathbb{R}^d$, and*

$$\eta_2 := -\limsup_{|x|\to\infty} \langle n(x), \nabla \log \pi(x) \rangle > 0. \tag{18}$$

Remark 13. *There exists $\beta > 0$ such that for x sufficiently large,*

$$\langle n(x), \nabla \log \pi(x) \rangle = \langle n(x), m(x) \rangle |\nabla \log \pi(x)| \leq -\beta.$$

Further, if $0 < -\langle n(x), m(x) \rangle \leq 1$, then $|\nabla \log \pi(x)| \geq \beta$.

Define the *symmetric proposal density family* $\mathfrak{C} := \{q : q(x,y) = q(x-y) = q(y-x)\}$. Our ergodicity result for adaptive Metropolis algorithms is based on the following assumptions.

Assumption 1 (Target Regularity). *The target distribution is absolutely continuous w.r.t. Lebesgue measure μ_d with a density π bounded away from zero and infinity on compact sets, and $\sup\limits_{x\in\mathcal{X}} \pi(x) < \infty$.*

Assumption 2 (Target Strongly Decreasing). *The target density π has continuous first derivatives and satisfies*

$$\eta_1 := -\limsup_{|x|\mapsto\infty} \langle n(x), m(x) \rangle > 0. \tag{19}$$

Assumption 3 (Proposal Uniform Local Positivity). *Assume that $\{q_\gamma : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}$. There exist $\zeta > 0$ such that*

$$\iota := \inf_{\gamma\in\mathcal{Y}} \inf_{|z|\leq\zeta} q_\gamma(z) > 0. \tag{20}$$

Given $0 < p < q < \infty$, for $u \in S^{d-1}$ (S^{d-1} is the unit hypersphere in $\mathbb{R}^d$.) and $\theta > 0$, define

$$C_{p,q}(u,\theta) := \left\{ z = a\xi \mid p \leq a \leq q, \, \xi \in S^{d-1}, \, |\xi - u| < \theta/3 \right\}. \tag{21}$$

Assumption 4 (Proposal Moment Condition)**.** *Suppose the target density* π *is exponentially tailed and* $\{q_\gamma : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}$*. Under Assumptions 2, assume that there are* $\epsilon \in (0, \eta_1)$*,* $\beta \in (0, \eta_2)$*,* δ*, and* Δ *with* $0 < \frac{3}{\beta\epsilon} \le \delta < \Delta \le \infty$ *such that*

$$\inf_{(u,\gamma) \in S^{d-1} \times \mathcal{Y}} \int_{C_{\delta,\Delta}(u,\epsilon)} |z| \, q_\gamma(z) \mu_d(dz) > \frac{3(e+1)}{\beta\epsilon(e-1)}. \tag{22}$$

Remark 14. *Under Assumption 3, let* $\tilde{P}(x, dy)$ *be the transition kernel of Metropolis-Hastings algorithm with the proposal distribution* $\tilde{Q}(x, \cdot) \sim \mathrm{Unif}(\bar{\mathrm{B}}^d(x, \zeta/2))$*. For any* $\gamma \in \mathcal{Y}$*,* $P_\gamma(x, dy) \ge \iota \, \mathrm{Vol}(\bar{\mathrm{B}}^d(0, \zeta/2)) \tilde{P}(x, dy)$*. Under Assumption 1, by [25, Theorem 2.2], any compact set is a small set for* $\tilde{P}$ *so that any compact set is a uniform small set for all* P_γ*.*

Remark 15. 1. *Assumption 4 means that the proposal family has uniform lower bound of the first moment on some local cone around the origin. The condition specifies that the tails of all proposal distributions can not be too light, and the quantity of the lower bound is given and dependent on the tail-decaying rate* η_2 *and the strongly decreasing rate* η_1 *of target distribution. Assumptions 1-4 are used to check S.G.E. which is just sufficient to Containment.*
2. *If the proposal distribution in* $\{q_\gamma : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}$ *is a mixture distribution with one fixed part, then Assumption 4 is relatively easy to check, because the integral in Eq. (22) can be estimated by the fixed part distribution. Especially for the lighter-than-exponentially tailed target, Assumption 4 can be reduced for this case. We will give a sufficient condition for Assumption 4 which can be applied to more general case, see Lemma 1.*

Now, we consider a particular class of target densities with tails which are heavier than exponential tails. It was previously shown by [12] that the Metropolis algorithm converges at any polynomial rate when proposal distribution is compact supported and the log density decreases hyperbolically at infinity, $\log \pi(x) \sim -|x|^s$, for $0 < s < 1$, as $|x| \to \infty$.

Definition 3 (Hyperbolic tail)**.** *The density function* $\pi(\cdot)$ *is twice continuously differentiable, and there exist* $0 < m < 1$ *and some finite positive constants* d_i, D_i*,* $i = 1, 2$ *such that for large enough* $|x|$*,*

$0 < d_0 |x|^m \le -\log \pi(x) \le D_0 |x|^m$;
$0 < d_1 |x|^{m-1} \le |\nabla \log \pi(x)| \le D_1 |x|^{m-1}$;
$0 < d_2 |x|^{m-2} \le \left|\nabla^2 \log \pi(x)\right| \le D_2 |x|^{m-2}$.

Assumption 5 (Proposal's Uniform Compact Support)**.** *Under Assumption 3, there exists some* $M > \zeta$ *such that all* $q_\gamma(\cdot)$ *with* $\gamma \in \mathcal{Y}$ *are supported entirely on* $\bar{\mathrm{B}}^d(0, M)$*.*

Theorem 6. *An adaptive Metropolis algorithm with Diminishing Adaptation is ergodic, under any of the following conditions:*
(i)*. The target density* π *is lighter-than-exponentially tailed, and Assumptions 1 – 3;*
(ii)*. The target density* π *is exponentially tailed, and Assumptions 1 – 4;*
(iii)*. The target density* π *is hyperbolically tailed, and Assumptions 1 – 3 and 5.*

For a proof of Theorem 6, see Section 5.7.

4.2 Specific Cases of Adaptive Metropolis-Hastings Algorithms

Here we discuss two specific cases of adaptations of Metropolis-Hastings algorithms. The first one (Example 3) is from [24] where the proposal density is a fixed distribution of two multivariate normal distributions, one with fixed small variance, another using the estimate of empirical covariance matrix from historical information as its variance. It is a slight variant of the original adaptive Metropolis algorithm of Haario et al. [14]. In the example, the target density has lighter-than-exponential tails. The second (Example 4) concerns with target densities with truly exponential tails.

Example 3. *Consider a d-dimensional target distribution $\pi(\cdot)$ on $\mathbb{R}^d$ satisfying Assumptions 1 - 2. We perform a Metropolis algorithm with proposal distribution given at the n^{th} iteration by $Q_n(x,\cdot) = N(x,(0.1)^2 I_d/d)$ for $n \leq 2d$; For $n > 2d$,*

$$Q_n(x,\cdot) = \begin{cases} (1-\theta)N(x,(2.38)^2\Sigma_n/d) + \theta N(x,(0.1)^2 I_d/d), & \Sigma_n \text{ is positive definite,} \\ N(x,(0.1)^2 I_d/d), & \Sigma_n \text{ is not positive definite,} \end{cases} \tag{23}$$

for some fixed $\theta \in (0,1)$, I_d is $d \times d$ identity matrix, and the empirical covariance matrix

$$\Sigma_n = \frac{1}{n}\left(\sum_{i=0}^{n} X_i X_i^\top - (n+1)\overline{X}_n \overline{X}_n^\top\right), \tag{24}$$

where $\overline{X}_n = \frac{1}{n+1}\sum_{i=0}^{n} X_i$, is the current modified empirical estimate of the covariance structure of the target distribution based on the run so far.

Remark 16. *The fixed part $N(x,(0.1)^2 I_d/d)$ can be replaced by $\mathrm{Unif}(\mathrm{B}^d(x,\tau))$ for some $\tau > 0$. For targets with lighter-than-exponential tails, τ can be an arbitrary positive value, because Assumption 3 holds. For targets with exponential tails, τ is dependent on η_1 and η_2.*

Remark 17. *The proposal $N(x,(2.38)^2\Sigma/d)$ is optimal in a particular large-dimensional context, [see 26, 22]. Thus the proposal $N(x,(2.38)^2\Sigma_n/d)$ is an effort to approximate this.*

Remark 18. *Commonly, the iterative form of Eq. (24) is more useful,*

$$\Sigma_n = \frac{n-1}{n}\Sigma_{n-1} + \frac{1}{n+1}\left(X_n - \bar{X}_{n-1}\right)\left(X_n - \bar{X}_{n-1}\right)^\top. \tag{25}$$

Proposition 5. *Suppose that the target density π is exponentially tailed. Under Assumptions 1-4, $\left|\overline{X}_n - \overline{X}_{n-1}\right|$ and $\|\Sigma_n - \Sigma_{n-1}\|_M$ converge to zero in probability where where $\|\cdot\|_M$ is matrix norm.*

For a proof of Proposition 5, see Section 5.8.

Theorem 7. *Suppose that the target density π in Example 3 is lighter-than-exponentially tailed. The algorithm in Example 3 is ergodic.*

Proof: Obviously, the proposal densities are uniformly bounded below. By Theorem 6 and Proposition 5, the adaptive Metropolis algorithm is ergodic. □

The following lemma can be used to check Assumption 4.

Lemma 1. *Suppose that the target density π is exponentially tailed and the proposal density family $\{q_\gamma : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}$. Suppose further that there is a function $q^-(z) := g(|z|)$, $q^- : \mathbb{R}^d \to \mathbb{R}^+$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$, some constants $M \geq 0$, $\epsilon \in (0,\eta_1)$, $\beta \in (0,\eta_2)$ and $\frac{3}{\beta\epsilon} \vee M < \delta < \Delta$ such that for $|z| \geq M$ with the property that $q_\gamma(z) \geq q^-(z)$ for $\gamma \in \mathcal{Y}$ and*

$$\frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right)\int_\delta^\Delta g(t)t^d dt > \frac{3(e+1)}{\beta\epsilon(e-1)}, \tag{26}$$

where η_1 is defined in Eq. (18), η_2 is defined in Eq. (19), $r := \frac{\epsilon}{18}\sqrt{36-\epsilon^2}$, and the incomplete beta function $\mathrm{Be}_x(t_1,t_2) := \int_0^x t^{t_1-1}(1-t)^{t_2-1}dt$, then Assumption 4 holds.

For a proof of Lemma 1, see Section 5.9.

We now consider a specific example to illustrate the theorem.

Example 4. *Consider the standard multivariate exponential distribution* $\pi(x) = c\exp(-\lambda\,|x|)$ *on* $\mathbb{R}^d$ *where* $\lambda > 0$. *We perform a Metropolis algorithm with proposal distribution in the family* $\{Q_\gamma(\cdot)\}_{\gamma\in\mathcal{Y}}$ *at the* n^{th} *iteration where*

$$Q_n(x,\cdot) = \begin{cases} \text{Unif}\left(\mathrm{B}^d(x,\Delta)\right), & n \le 2d, \text{ or } \Sigma_n \text{ is nonsingular,} \\ (1-\theta)N(x,(2.38)^2\Sigma_n/d) + \theta\,\text{Unif}\left(\mathrm{B}^d(x,\Delta)\right), & n > 2d, \text{ and } \Sigma_n \text{ is singular,} \end{cases} \tag{27}$$

for $\theta \in (0,1)$, $\text{Unif}\left(\mathrm{B}^d(x,\Delta)\right)$ *is a uniform distribution on the hyperball* $\mathrm{B}^d(x,\Delta)$ *with the center* x *and the radius* Δ, *and* Σ_n *is as defined in Eq. (24).*

Proposition 6. *There exists a large enough* $\Delta > 0$ *such that the adaptive Metropolis algorithm of Example 4 is ergodic.*

For a proof of Proposition 6, see Section 5.10.

Remark 19. *Concurrent with our research, Saksman and Vihola [30] recently proved some related results about the original Adaptive Metropolis (AM) algorithm of [14], assuming lighter-than-exponential tails of the target distribution as in our Theorem 7 above. Their Theorem 13 shows that if the target density is regular, strongly decreasing, and strongly lighter-than-exponentially tailed (i.e.,* $\limsup_{|x|\to\infty} \frac{\langle n(x), \nabla \log \pi(x)\rangle}{|x|^{\rho-1}} = -\infty$ *for some* $\rho > 1$*), then strong laws of large numbers and central limit theorems hold in the adaptive setting.*

5 Proofs of the Results

5.1 Proofs Related to Example 1

Proof of Proposition 1: Since the adaptation is state-independent, the stationarity is preserved. So, the adaptive MCMC $X_n \sim \delta P_{\theta_0}P_{\theta_1}P_{\theta_2}\cdots P_{\theta_{n-1}}(\cdot)$ for $n \ge 0$ where $\delta := (\delta^{(1)}, \delta^{(2)})$ is the initial distribution.

The part (i). Consider $\left\|P_{\theta_{n+1}}(x,\cdot) - P_{\theta_n}(x,\cdot)\right\|_{\text{TV}}$. For any $x \in \mathcal{X}$,

$$\left\|P_{\theta_{n+1}}(x,\cdot) - P_{\theta_n}(x,\cdot)\right\|_{\text{TV}} = |\theta_{n+1} - \theta_n| \to 0.$$

Thus, for $r > 0$ Diminishing Adaptation holds.

By some algebra,

$$\|P^n_\theta(x,\cdot) - \pi(\cdot)\|_{\text{TV}} = \frac{1}{2}\,|1-2\theta|^n\,. \tag{28}$$

Hence, for any $\epsilon > 0$,

$$M_\epsilon(X_n,\theta_n) \ge \frac{\log(\epsilon) - \log(1/2)}{\log|1-2\theta_n|} \to +\infty \qquad \text{as } n \to \infty. \tag{29}$$

Therefore, the stochastic process $\{M_\epsilon(X_n,\theta_n) : n \ge 0\}$ is not bounded in probability.

The parts (ii) and (iii). Let $\mu_n := \left(\mu_n^{(1)}, \mu_n^{(2)}\right) := \delta P_{\theta_0}\cdots P_{\theta_n}$. So,

$$\mu^{(1)}_{n+1} = \mu^{(1)}_n - \theta_{n+1}\left(\mu^{(1)}_n - \mu^{(2)}_n\right) \text{ and } \mu^{(2)}_{n+1} = \mu^{(2)}_n + \theta_{n+1}\left(\mu^{(1)}_n - \mu^{(2)}_n\right).$$

Hence,

$$\mu_{n+1}^{(1)} - \mu_{n+1}^{(2)} = \left(\delta^{(1)} - \delta^{(2)}\right) \prod_{k=0}^{n+1} (1 - 2\theta_k).$$

For $r > 1$, $\prod_{k=0}^{n+1}(1 - 2\theta_k)$ converges to some $\alpha \in (0, 1)$ as n goes to infinity. $\mu_{n+1}^{(1)} - \mu_{n+1}^{(2)} \to \left(\delta^{(1)} - \delta^{(2)}\right)\alpha$. For $0 < r \le 1$, $\mu_{n+1}^{(1)} - \mu_{n+1}^{(2)} \to 0$. Therefore, for $r > 1$ ergodicity to Uniform distribution does not hold, and for $0 < r \le 1$ ergodicity holds. $\square$

Proof of Proposition 2: From Eq. (28), for $\epsilon > 0$, $M_\epsilon(X_{2k-1}, \theta_{2k-1}) \ge \frac{\log(\epsilon) - \log(1/2)}{\log|1-1/k|} \to \infty$ as $k \to \infty$. So, Containment does not hold.
$\left\|P_{\theta_{2k}}(x, \cdot) - P_{\theta_{2k-1}}(x, \cdot)\right\|_{\mathrm{TV}} = \left|\frac{1}{2} - \frac{1}{2k}\right| \to \frac{1}{2}$ as $k \to \infty$. So Diminishing Adaptation does not hold.
Let $\delta := (\delta^{(1)}, \delta^{(2)})$ be the initial distribution and $\mu_n := (\mu_n^{(1)}, \mu_n^{(2)}) = \delta P_{\theta_0} \cdots P_{\theta_n}$. $\mu_n^{(1)} - \mu_n^{(2)} = (\delta^{(1)} - \delta^{(2)}) 2^{-[n/2]-1} \prod_{k=1}^{[(n+1)/2]} \left(1 - \frac{1}{2k}\right) \to 0$ as n goes to infinity. So ergodicity holds. $\square$

5.2 Proofs of Theorem 2 and Corollary 1

Proof of Theorem 2: Fix $\epsilon > 0$. For any $\delta > 0$, taking $K > 0$ such that $\pi(\mathcal{D}_K^c) < \delta/2$. For the set $\mathcal{D}_K$, there exists M such that

$$\sup_{\mathcal{D}_K \times \mathcal{Y}} \left\|P_\gamma^M(x, \cdot) - \pi(\cdot)\right\|_{\mathrm{TV}} < \epsilon.$$

Hence, for any $(x_0, \gamma_0) \in \mathcal{X} \times \mathcal{Y}$, by the ergodicity of the adaptive MCMC $\{X_n\}_n$, there exists some $N > 0$ such that $n > N$,

$$\left|\mathbb{P}_{(x_0,\gamma_0)}(X_n \in \mathcal{D}_K^c) - \pi(\mathcal{D}_K^c)\right| < \delta/2.$$

So, for $(X_n, \Gamma_n) \in (\mathcal{D}_K, \mathcal{Y})$,

$$[X_n \in \mathcal{D}_K] = [(X_n, \Gamma_n) \in \mathcal{D}_K \times \mathcal{Y}] \subset [M_\epsilon(X_n, \Gamma_n) \le M].$$

Hence,

$$\begin{aligned} & \mathbb{P}_{(x_0,\gamma_0)}(M_\epsilon(X_n, \Gamma_n) > M) \\ \le\ & \mathbb{P}_{(x_0,\gamma_0)}((X_n, \Gamma_n) \in (\mathcal{D}_K \times \mathcal{Y})^c) \\ =\ & \mathbb{P}_{(x_0,\gamma_0)}(X_n \in \mathcal{D}_K^c) \\ \le\ & \left|\mathbb{P}_{(x_0,\gamma_0)}(X_n \in \mathcal{D}_K^c) - \pi(\mathcal{D}_K^c)\right| + \pi(\mathcal{D}_K^c) < \delta. \end{aligned}$$

Therefore, Containment holds. $\square$

Proof of Corollary 1: Using the same technique in Theorem 2, for large enough $M > 0$,

$$\begin{aligned} & \mathbb{P}_{(x_0,\gamma_0)}(M_\epsilon(X_n, \Gamma_n) > M) \\ \le\ & \mathbb{P}_{(x_0,\gamma_0)}((X_n, \Gamma_n) \in (D_k \times \mathcal{Y}_k)^c) \\ \le\ & \mathbb{P}_{(x_0,\gamma_0)}(X_n \in D_k^c) + \mathbb{P}_{(x_0,\gamma_0)}(\Gamma_n \in \mathcal{Y}_k^c) \\ \le\ & \left|\mathbb{P}_{(x_0,\gamma_0)}(X_n \in \mathcal{D}_K^c) - \pi(\mathcal{D}_K^c)\right| + \pi(\mathcal{D}_K^c) + \mathbb{P}_{(x_0,\gamma_0)}(\Gamma_n \in \mathcal{Y}_k^c). \end{aligned}$$

Since $\{\Gamma_n : n \ge 0\}$ is bounded in probability, the result holds. $\square$

5.3 Proof of Proposition 3

First, we show that Diminishing Adaptation holds.

Lemma 2. *For the adaptive chain $\{X_n : n \geq 0\}$ defined in Example 2, the adaptation is diminishing.*

Proof: For $\gamma = 1$, obviously the proposal density is $q_\gamma(x, y) = \varphi(y - x)$ where $\varphi(\cdot)$ is the density function of standard normal distribution. For $\gamma = -1$, the random variable $1/x + Z_n$ has the density $\varphi(y - 1/x)$ so the random variable $1/(1/x + Z_n)$ has the density $q_\gamma(x, y) = \varphi(1/y - 1/x)/y^2$.

The proposal density

$$q_\gamma(x, y) = \begin{cases} \varphi(y - x) & \gamma = 1 \\ \varphi(1/y - 1/x)/y^2 & \gamma = -1 \end{cases}$$

For $\gamma = 1$, the acceptance rate is $\min\left(1, \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}\right)\mathbb{I}(y \in \mathcal{X}) = \frac{1+x^2}{1+y^2}\mathbb{I}(y > 0)$. For $\gamma = -1$, the acceptance rate is $\min\left(1, \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}\right)\mathbb{I}(y \in \mathcal{X}) = \min\left(1, \frac{\frac{1}{1+y^2}\varphi(1/x-1/y)/x^2}{\frac{1}{1+x^2}\varphi(1/y-1/x)/y^2}\right)\mathbb{I}(y > 0) = \min\left(1, \frac{1+x^{-2}}{1+y^{-2}}\right)\mathbb{I}(y > 0)$.

So for $\gamma \in \mathcal{Y}$, the acceptance rate is

$$\alpha_\gamma(x, y) := \min\left(1, \frac{\pi(y)q_\gamma(y, x)}{\pi(x)q_\gamma(x, y)}\right)\mathbb{I}(y \in \mathcal{X}) = \min\left(1, \frac{1 + x^{2\gamma}}{1 + y^{2\gamma}}\right)\mathbb{I}(y \in \mathcal{X}). \tag{30}$$

From Eq. (4), $[\Gamma_n \neq \Gamma_{n-1}] = [X_n^{\Gamma_{n-1}} < 1/n]$. Since the joint process $\{(X_n, \Gamma_n) : n \geq 0\}$ is a time inhomogeneous Markov chain,

$$\begin{aligned}
\mathbb{P}(\Gamma_n \neq \Gamma_{n-1}) &= \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{P}(X_n^{\Gamma_{n-1}} < 1/n \mid X_{n-1} = x, \Gamma_{n-1} = \gamma)\mathbb{P}(X_{n-1} \in dx, \Gamma_{n-1} \in d\gamma) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} P_\gamma(x, [t > 0 : t^\gamma < 1/n])\mathbb{P}(X_{n-1} \in dx, \Gamma_{n-1} \in d\gamma) \\
&= \int_{[x^\gamma \geq 1/(n-1)]} P_\gamma(x, [t > 0 : t^\gamma < 1/n])\mathbb{P}(X_{n-1} \in dx, \Gamma_{n-1} \in d\gamma)
\end{aligned}$$

where the second equality is from Eq. (1), and the last equality is from $\mathbb{P}(X_n^{\Gamma_n} \geq 1/n) = 1$ implied by Eq. (4).

So for any $(x, \gamma) \in [(t, s) \in \mathcal{X} \times \mathcal{Y} : t^s \geq 1/(n-1)]$,

$$P_\gamma(x, [t > 0 : t^\gamma < 1/n]) = \int_0^\infty \mathbb{I}(y^\gamma < 1/n)q_\gamma(x, y)dy = \int_{-x^\gamma}^{-x^\gamma+1/n} \varphi(z)dz.$$

Since $-x^\gamma + 1/(n-1) < 0$,

$$\frac{1}{n}\varphi(-x^\gamma) \leq P_\gamma(x, [t > 0 : t^\gamma < 1/n]) \leq \frac{\varphi(0)}{n}. \tag{31}$$

We have that

$$\mathbb{P}(\Gamma_n \neq \Gamma_{n-1}) \leq \frac{1}{\sqrt{2\pi}n}. \tag{32}$$

Therefore, for any $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}} \left\|P_{\Gamma_n}(x,\cdot) - P_{\Gamma_{n-1}}(x,\cdot)\right\|_{\mathrm{TV}} > \epsilon\right) \leq \mathbb{P}(\Gamma_n \neq \Gamma_{n-1}) \to 0.$$

$\square$

From Eq. (30), at the n^{th} iteration, the acceptance rate is $\alpha_{\Gamma_{n-1}}(X_{n-1}, Y_n) = \min\left(1, \frac{1+X_{n-1}^{2\Gamma_{n-1}}}{1+Y_n^{2\Gamma_{n-1}}}\right)\mathbb{I}(Y_n > 0)$. Let us denote $\tilde{Y}_n := Y_n^{\Gamma_{n-1}}$ and $\tilde{X}_n := X_n^{\Gamma_n}$. The acceptance rate is equal to

$$\min\left(1, \frac{1+\tilde{X}_{n-1}^2}{1+\tilde{Y}_n^2}\right)\mathbb{I}(\tilde{Y}_n > 0).$$

From Eq. (4), $X_n^{\Gamma_n} = X_n^{-\Gamma_{n-1}}\mathbb{I}(X_n^{\Gamma_{n-1}} < 1/n) + X_n^{\Gamma_{n-1}}\mathbb{I}(X_n^{\Gamma_{n-1}} \geq 1/n)$. When Y_n is accepted, i.e. $X_n = Y_n$,

$$[\tilde{Y}_n < 1/n] = [X_n^{\Gamma_{n-1}} < 1/n] \text{ and } X_n^{\Gamma_n} = \tilde{Y}_n^{-1}\mathbb{I}(\tilde{Y}_n < 1/n) + \tilde{Y}_n\mathbb{I}(\tilde{Y}_n \geq 1/n).$$

On the other hand, from Eq. (3), the conditional distribution $\tilde{Y}_n \mid \tilde{X}_{n-1}$ is $N(\tilde{X}_{n-1}, 1)$.

From the above discussion, the chain $\tilde{\mathbf{X}} := \{\tilde{X}_n : n \geq 0\}$ can be constructed according to the following procedure. Define the independent random variables $Z_n \overset{\text{iid}}{\sim} N(0,1)$, $U_n \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$, and $T_n \overset{\text{iid}}{\sim} \text{Unif}(0,1)$.

Let $\tilde{X}_0 = X_0^{\Gamma_0}$. At each time $n \geq 1$, define the variable

$$\tilde{Y}_n := \tilde{X}_{n-1} - U_n\left|Z_n\right| + (1-U_n)\left|Z_n\right|. \tag{33}$$

Clearly, $-U_n\left|Z_n\right| + (1-U_n)\left|Z_n\right| \overset{\text{d}}{=} N(0,1)$ ($\overset{\text{d}}{=}$ means equal in distribution).

If $T_n < \min\left(1, \frac{1+\tilde{X}_{n-1}^2}{1+\tilde{Y}_n^2}\right)\mathbb{I}(\tilde{Y}_n > 0)$ then

$$\tilde{X}_n = \mathbb{I}(\tilde{Y}_n < 1/n)\tilde{Y}_n^{-1} + \mathbb{I}(\tilde{Y}_n \geq 1/n)\tilde{Y}_n; \tag{34}$$

otherwise $\tilde{X}_n = \tilde{X}_{n-1}$.

Note that:
1. The process $\tilde{\mathbf{X}}$ is a time inhomogeneous Markov chain.
2. $\mathbb{P}(\tilde{X}_n \geq 1/n) = 1$ for $n \geq 1$.
3. At the time n, U_n indicates the proposal direction ($U_n = 0$: try to jump towards infinity; $U_n = 1$: try to jump towards zero). $|Z_n|$ specifies the step size if the proposal value Y_n is accepted. T_n is used to check whether the proposal value Y_n is accepted or not. When $U_n = 1$ and $\tilde{Y}_n > 0$, Eq. (34) is always run.

For two integers $0 \leq s \leq t$ and a process X and a set $A \subset \mathcal{X}$, denote $[X_{s:t} \in A] := [X_s \in A; X_{s+1} \in A; \cdots; X_t \in A]$ and $s:t := \{s, s+1, \cdots, t\}$. For a value $x \in \mathbb{R}$, denote the largest integer less than x by $[x]$.

In the following proofs for the example, we use the notation in the procedure of constructing the process $\tilde{\mathbf{X}}$.

Lemma 3. *Let $a = \left(\frac{1}{2} - \frac{7\sqrt{2}}{12\sqrt{\pi}}\right)^{-2}$. Given $0 < r < 1$, for $[x] > 12^{\frac{1}{1-r}}$*

$$\mathbb{P}\left(\exists i \in (k+1) : (k + [x]^{1+r}),\ \tilde{X}_i < x/2 \mid \tilde{X}_k = x\right) \leq \frac{[x]^{1+r}}{\left(\frac{[x]}{2} - \frac{7\sqrt{2}[x]^r}{\sqrt{\pi}}\right)^2} \leq \frac{a}{[x]^{1-r}}.$$

Proof: The process $\tilde{\mathbf{X}}$ is generated through the underlying processes $\{(\tilde{Y}_j, Z_j, U_j, T_j) : j \geq 1\}$ defined in Eq. (33) – Eq. (34). Conditional on $[\tilde{X}_k = x]$, we can construct an auxiliary chain $\mathbf{B} := \{B_j : j \geq k\}$ that behaves like an asymmetric random walk until $\tilde{\mathbf{X}}$ reaches below $x/2$, and $\mathbf{B}$ is always dominated from above by $\tilde{\mathbf{X}}$.

It is defined as that $B_k = \tilde{X}_k$; For $j > k$, if $\tilde{X}_{j-1} < x/2$ then $B_j := \tilde{X}_j$, otherwise

1. If proposing towards zero ($U_j = 1$) then $\mathbf{B}$ also jumps in the same direction with the step size $|Z_j|$ (in this case, the acceptance rate $\min\left(1, \frac{1+\tilde{X}_{j-1}^2}{1+\tilde{Y}_j^2}\right)$ is equal to 1);
2. If proposing towards infinity ($U_j = 0$), then B_j is assigned the value $B_{j-1} + |Z_j|$ (the jumping direction of $\mathbf{B}$ at the time j is same as $\tilde{\mathbf{X}}$) with the acceptance rate $\frac{1+(x/2)^2}{1+(x/2+|Z_j|)^2}$ (independent of $\tilde{X}_{j-1}$), i.e. for $j > k$,

$$B_j := \mathbb{I}(\tilde{X}_{j-1} < x/2)\tilde{X}_j + \mathbb{I}(\tilde{X}_{j-1} \geq x/2)\left(B_{j-1} - I_j(x)\right) \tag{35}$$

where

$$I_j(x) := U_j\,|Z_j| - (1 - U_j)\,|Z_j|\,\mathbb{I}\left(T_j < \frac{1+(x/2)^2}{1+(x/2+|Z_j|)^2}\right). \tag{36}$$

Note that

1. $\{Z_j, U_j, T_j : j > k\}$ are independent so $\{I_j(x) : j > k\}$ are independent.
2. When $\tilde{X}_{j-1} > x/2$ and $U_j = 0$ (proposing towards infinity), the acceptance rate $1 > \frac{1+\tilde{X}_{j-1}^2}{1+\tilde{Y}_j^2} \geq \frac{1+(x/2)^2}{1+(x/2+|Z_j|)^2}$, so that $\left[T_j < \frac{1+(x/2)^2}{1+(x/2+|Z_j|)^2}\right] \subset \left[T_j < \frac{1+\tilde{X}_{j-1}^2}{1+\tilde{Y}_j^2}\right]$ which is equivalent to $[B_j - B_{j-1} = |Z_j|] \subset [\tilde{X}_j - \tilde{X}_{j-1} = |Z_j|]$. Therefore, $\mathbf{B}$ is always dominated from above by $\tilde{\mathbf{X}}$.

Conditional on $[\tilde{X}_k = x]$,

$$[\exists i \in (k+1):(k+[x]^{1+r}),\ \tilde{X}_i < x/2] \subset [\exists i \in (k+1):(k+[x]^{1+r}),\ B_i < x/2]$$

and for $i \in (k+1):(k+[x]^{1+r})$,

$$\begin{aligned}
&[B_{k:(i-1)} \geq x/2; B_i < x/2]\\
\subset&[B_k \geq x/2; B_k - \sum_{l=k+1}^{t-1} I_l(x) \geq x/2 \text{ for all } t \in (k+1):i; B_k - \sum_{l=k+1}^{i} I_l(x) < x/2].
\end{aligned}$$

So,

$$\begin{aligned}
&\mathbb{P}\left(\exists i \in (k+1):(k+[x]^{1+r}),\ \tilde{X}_i < x/2 \mid \tilde{X}_k = x\right)\\
\leq&\mathbb{P}\left(\exists i \in (k+1):(k+[x]^{1+r}),\ B_k - \sum_{j=k+1}^{i} I_j(x) < x/2 \mid B_k = x\right)\\
\leq&\mathbb{P}(\max_{l \in 1:[x]^{1+r}} \tilde{S}_l > x/2)\\
=&\mathbb{P}(\max_{l \in 1:q} \tilde{S}_l > q^{1/(1+r)}/2)
\end{aligned}$$

where $\tilde{S}_0 = 0$ and $\tilde{S}_l = \sum_{j=1}^{l} I_{k+j}(x)$ and $q = [x]^{1+r}$. $\{I_j(x) : k < j \leq k+l\}$ and B_k are independent so that the right hand side of the above equation is independent of k.

By some algebra,

$$0 \le \mathbb{E}[I_i(x)] = \frac{1}{2}\mathbb{E}\left[\frac{|Z_i|^2 (x + |Z_i|)}{1 + (x/2 + |Z_i|)^2}\right] \le \frac{2}{x}\mathbb{E}\left[|Z_i|^2 (1 + |Z_i|)\right] < \frac{7\sqrt{2}}{\sqrt{\pi}x},$$

$$\mathrm{Var}[I_i(x)] = \frac{1}{2} + \frac{1}{2}\mathbb{E}\left[|Z_i|^2 \frac{1 + (x/2)^2}{1 + (x/2 + |Z_i|)^2}\right] - \frac{1}{4}\left(\mathbb{E}\left[\frac{|Z_i|^2 (x + |Z_i|)}{1 + (x/2 + |Z_i|)^2}\right]\right)^2 \in [0, 1].$$

Let $\mu_l = \mathbb{E}[\tilde{S}_l]$ and $S_l = \tilde{S}_l - \mu_l$ and note that μ_l is increasing as l increases, and $\mu_q \in [0, \frac{7\sqrt{2}q}{\sqrt{\pi}}]$. So $\{S_i : i = 1, \cdots, q\}$ is a Martingale. By Kolmogorov Maximal Inequality,

$$\begin{aligned}
\mathbb{P}(\max_{l \in 1:q} \tilde{S}_l > q^{1/(1+r)}/2) \le & \mathbb{P}(\max_{l \in 1:q} S_l > q^{1/(1+r)}/2 - \mu_q) \\
\le & \frac{q\mathrm{Var}[I_k(x)]}{(q^{1/(1+r)}/2 - \mu_q)^2} \\
\le & \frac{[x]^{1+r}}{\left(\frac{[x]}{2} - \frac{7\sqrt{2}[x]^r}{\sqrt{\pi}}\right)^2} < \frac{a}{[x]^{1-r}}.
\end{aligned}$$

The last second inequality is from $[x] > 12^{\frac{1}{1-r}} > \left(\frac{14\sqrt{2}}{\sqrt{\pi}}\right)^{\frac{1}{1-r}}$ implying $\frac{[x]}{2} > \frac{7\sqrt{2}[x]^r}{\sqrt{\pi}}$. □

Assume that X_n converges weakly to $\pi(\cdot)$. Take some $c > 1$ such that for the set $D = (1/c, c)$, $\pi(D) = 9/10$. Taking a $r \in (0, 1)$, there exists $N > 2c \vee 12^{\frac{1}{1-r}} \vee \frac{a}{0.5}^{\frac{1}{1-r}} \vee 2^{1/r} \exp(\frac{1}{0.8\varphi(-c)r})$ (a is defined in Lemma 3) such that for any $n > N + 1$, $\mathbb{P}(X_n \in D) > 0.8$. Since $[X_n \in D] = [X_n^{\Gamma_n} \in D]$ and $\mathbf{X}^{\mathbf{\Gamma}} \overset{\mathrm{d}}{=} \tilde{\mathbf{X}}$, $\mathbb{P}(\tilde{X}_n \in D) > 0.8$. So, $\mathbb{P}(\tilde{X}_n > \frac{n}{2}) < 0.2$ for $n > N$.

Let $m = \exp(\frac{1}{0.8\varphi(-c)})(n+1) - 1$ that implies $m > n$, $m - n < n^{1+r}$ (because $n > 2^{1/r} \exp(\frac{1}{0.8\varphi(-c)r})$), and $\log(\frac{m+1}{n+1}) = \frac{1}{0.8\varphi(-c)}$. Then

$$0.2 > \mathbb{P}(\tilde{X}_m > \frac{n}{2}) \ge \sum_{j=n}^{m-1} \mathbb{P}(\tilde{X}_j \in D; \tilde{Y}_{j+1} < \frac{1}{j+1}; \tilde{X}_{(j+1):m} > \frac{n}{2}). \tag{37}$$

From Eq. (33) and Eq. (34), $[\tilde{Y}_{i+1} < \frac{1}{i+1}] = [\tilde{X}_{i+1} = \frac{1}{\tilde{Y}_{i+1}} > i + 1]$ for any $i > 1$. Consider $j \in n : (m-1)$. Since $\tilde{\mathbf{X}}$ is a time inhomogeneous Markov chain,

$$\begin{aligned}
& \mathbb{P}\left(\tilde{X}_j \in D; \tilde{Y}_{j+1} < \frac{1}{j+1}; \tilde{X}_{(j+1):m} > n/2\right) \\
= & \mathbb{P}(\tilde{X}_j \in D)\mathbb{P}\left(\tilde{X}_{j+1} = \tilde{Y}_{j+1} < \frac{1}{j+1} \mid \tilde{X}_j \in D\right) \\
& \mathbb{P}\left(\tilde{X}_{(j+2):m} > \frac{n}{2} \mid \tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j + 1\right) \\
= & \mathbb{P}(\tilde{X}_j \in D)\mathbb{P}\left(\tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j + 1 \mid \tilde{X}_j \in D\right) \\
& \left(1 - \mathbb{P}\left(\tilde{X}_t \le n/2 \text{ for some } t \in (j+1) : m \mid \tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j + 1\right)\right).
\end{aligned}$$

From Eq. (31), for any $x \in D$,

$$\mathbb{P}(\tilde{Y}_{j+1} < \frac{1}{j+1} \mid \tilde{X}_j = x) = P_1(x, \{t \in \mathcal{X} : t < 1/(j+1)\}) \in \left[\frac{\varphi(-c)}{j+1}, \frac{\varphi(0)}{j+1}\right].$$

So,

$$\mathbb{P}(\tilde{Y}_{j+1} < \frac{1}{j+1} \mid \tilde{X}_j \in D) \geq \frac{\varphi(-c)}{j+1}.$$

Hence, for $x > j+1$,

$$\begin{aligned}
&\mathbb{P}\left(\tilde{X}_t \leq n/2 \text{ for some } t \in (j+1):m \mid \tilde{X}_{j+1} = x\right) \\
\leq &\mathbb{P}\left(\tilde{X}_t \leq x/2 \text{ for some } t \in (j+1):m \mid \tilde{X}_{j+1} = x\right) \\
\leq &\mathbb{P}\left(\tilde{X}_t \leq x/2 \text{ for some } t \in (j+1):(j+[x]^{1+r}) \mid \tilde{X}_{j+1} = x\right) \\
\leq &\frac{a}{[x]^{1-r}} \leq \frac{a}{n^{1-r}},
\end{aligned}$$

because of $x/2 > n/2$, $m - n < n^{1+r}$, and Lemma 3. Thus,

$$\mathbb{P}\left(\tilde{X}_t \leq n/2 \text{ for some } t \in (j+1):m \mid \tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j+1\right) \leq \frac{a}{n^{1-r}}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\tilde{X}_m > \frac{n}{2}) \geq &0.8\varphi(-c)(1 - \frac{a}{n^{1-r}}) \sum_{j=n}^{m-1} \frac{1}{j+1} \\
\geq &0.8\varphi(-c)(1 - \frac{a}{n^{1-r}}) \log((m+1)/(n+1)) = (1 - \frac{a}{n^{1-r}}) > 0.5.
\end{aligned}$$

Contradiction! By Lemma 2, Containment does not hold.

5.4 Proof of Proposition 4

Fix $x_0 \in \mathcal{X}$, $\gamma_0 \in \mathcal{Y}$. By the condition (iii) and the Borel-Cantelli Lemma, $\forall \epsilon > 0$, $\exists N_0(x_0, \gamma_0, \epsilon) > 0$ such that $\forall n \geq N_0$,

$$\mathbb{P}_{(x_0,\gamma_0)}(\Gamma_n = \Gamma_{n+1} = \cdots) > 1 - \epsilon/2. \tag{38}$$

Construct a new chain $\{\tilde{X}_n : n \geq 0\}$ which satisfies that for $n \leq N_0$, $\tilde{X}_n = X_n$, and for $n \geq N_0$, $\tilde{X}_n \sim P^{n-N_0}_{\Gamma_{N_0}}(\tilde{X}_{N_0}, \cdot)$. So, for any $n > N_0$ and any set $A \in \mathcal{B}(\mathcal{X})$, by the condition (ii),

$$\left|\mathbb{P}_{(x_0,\gamma_0)}(X_n \in A, \Gamma_{N_0} = \cdots = \Gamma_{n-1}) - \mathbb{P}_{(x_0,\gamma_0)}(\tilde{X}_n \in A)\right| \leq \epsilon/2.$$

Since the condition (i) holds, suppose that for some $K > 0$, $\mathcal{Y} = \{y_1, \cdots, y_K\}$. Denote $\mu_i(\cdot) = \mathbb{P}_{(x_0,\gamma_0)}(\tilde{X}_{N_0} \in \cdot \mid \Gamma_{N_0} = y_i)$ for $i = 1, \cdots, K$. Because of the condition (ii), for $n > N_0$,

$$
\begin{aligned}
&\mathbb{P}_{(x_0,\gamma_0)}(\tilde{X}_n \in A) \\
=&\sum_{i=1}^{K} \mathbb{P}_{(x_0,\gamma_0)}(\tilde{X}_n \in A, \Gamma_{N_0} = y_i) \\
=&\sum_{i=1}^{K} \int_{\mathcal{X}^{N_0} \cap [\gamma_{N_0} = y_i]} P_{\gamma_0}(x_0, dx_1) \cdots P_{\gamma_{N_0-1}}(x_{N_0-1}, dx_{N_0}) P_{y_i}^{n-N_0}(x_{N_0}, A) \\
=&\sum_{i=1}^{K} \mathbb{P}_{(x_0,\gamma_0)}(\Gamma_{N_0} = y_i) \mu_i P_{y_i}^{n-N_0}(A).
\end{aligned}
$$

By the condition (i), there exists $N_1(x_0, \gamma_0, \epsilon, N_0) > 0$ such that for $n > N_1$,

$$\sup_{i \in \{1,\cdots,K\}} \left\| \mu_i P_{y_i}^n(\cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} < \epsilon/2.$$

So, for any $n > N_0 + N_1$, any $A \in \mathcal{B}(\mathcal{X})$,

$$
\begin{aligned}
&\left| \mathbb{P}_{(x_0,\gamma_0)}(X_n \in A) - \pi(A) \right| \\
\leq &\left| \mathbb{P}_{(x_0,\gamma_0)}(X_n \in A) - \mathbb{P}_{(x_0,\gamma_0)}(\tilde{X}_n \in A) \right| + \\
&\left| \mathbb{P}_{(x_0,\gamma_0)}(\tilde{X}_n \in A) - \pi(A) \right| \\
\leq &(\epsilon/2 + \epsilon/2) + \epsilon/2 = 3\epsilon/2.
\end{aligned}
$$

Therefore, the adaptive MCMC $\{X_n : n \geq 0\}$ is ergodic. □

5.5 Proofs of Section 2.3

First we recall a previous result of [28, Theorem 5].

Proposition 7 ([28])**.** *Let $P(x, dy)$ be a Markov kernel on the state space $\mathcal{X}$. Suppose there is a set $C \subset \mathcal{X}$, $\delta > 0$, some integer $m > 0$, and a probability measure ν_m on $\mathcal{X}$ such that*

$$P^m(x, \cdot) \geq \delta \nu_m(\cdot) \text{ for } x \in C.$$

Suppose further that there exist $0 < \lambda < 1$, $b > 0$, and a function $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ such that

$$\mathbb{E}\left[h(X_1, Y_1) \mid X_0 = x, Y_0 = y\right] \leq \lambda h(x, y) + b \mathbb{I}_{C \times C}((x, y)).$$

Let $A := \sup_{(x,y) \in C \times C} \mathbb{E}[h(X_m, Y_m) \mid X_0 = x, Y_0 = y]$, $\mu := \mathcal{L}(X_0)$ be the initial distribution, and π be the stationary distribution. Then for any integer $j > 0$,

$$\|\mathcal{L}(X_n) - \pi\|_{\mathrm{TV}} \leq (1 - \delta)^{[j/m]} + \lambda^{n-jm+1} A^{j-1} \mathbb{E}_{\mu \times \pi}[h(X_0, Y_0)].$$

We now proceed to the proofs for this section.

Proof of Theorem 3: Let $\{X_n^{(\gamma)} : n \geq 0\}$ and $\{X_n^{(\gamma)} : n \geq 0\}$ be two realizations of P_γ for $\gamma \in \mathcal{Y}$. Define $h(x,y) := (V(x)+V(y))/2$. From (ii) of S.G.E., $\mathbb{E}[h(X_1^{(\gamma)}, Y_1^{(\gamma)}) \mid X_0^{(\gamma)} = x, Y_0^{(\gamma)} = y] \leq \lambda h(x,y) + b\mathbb{I}_{C\times C}((x,y))$. It is not difficult to get $P_\gamma^m V(x) \leq \lambda^m V(x) + bm$ so $A := \sup_{(x,y)\in C\times C} \mathbb{E}[h(X_m^{(\gamma)}, Y_m^{(\gamma)}) \mid X_0^{(\gamma)} = x, Y_0^{(\gamma)} = y] \leq \lambda^m \sup_C V + bm =: B$.

Consider $\mathcal{L}(X_0^{(\gamma)}) = \delta_x$ and $j := \sqrt{n}$. By Proposition 7,

$$\left\|P_\gamma^n(x,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}} \leq (1-\delta)^{[\sqrt{n}/m]} + \lambda^{n-\sqrt{n}m+1} B^{\sqrt{n}-1}(V(x)+\pi(V))/2. \tag{39}$$

Note that the quantitative bound is dependent on x, n, δ, m, C, V and π, and independent of γ. As n goes to infinity, the uniform quantitative bound of all $\left\|P_\gamma^n(x,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}}$ tends to zero for any $x \in \mathcal{X}$.

Let $\{X_n : n \geq 0\}$ be the adaptive MCMC satisfying S.G.E. From (ii) of S.G.E., $\sup_n \mathbb{E}[V(X_n) \mid X_0 = x, \Gamma_0 = \gamma_0] < \infty$ so the process $\{V(X_n) : n \geq 0\}$ is bounded in probability. Therefore, for any $\epsilon > 0$, $\{M_\epsilon(X_n, \Gamma_n) : n \geq 0\}$ is bounded in probability given any $X_0 = x$ and $\Gamma_0 = \gamma_0$. □

Proof of Corollary 2: From Eq. (6), letting $\lambda = \limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < 1$, there exists some positive constant K such that $\sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < \frac{\lambda+1}{2}$ for $|x| > K$. By $V > 1$, $P_\gamma V(x) < \frac{\lambda+1}{2} V(x)$ for $|x| > K$. $P_\gamma V(x) \leq \frac{\lambda+1}{2} V(x) + b\mathbb{I}_{\{z\in\mathcal{X}:|z|\leq K\}}(x)$ where $b = \sup_{x\in\{z\in\mathcal{X}:|z|\leq K\}} V(x)$. □

5.6 Proof of Theorem 5

The theorem follows from Theorem 8, Theorem 9, Theorem 10, and Lemma 4 below. Theorem 10 shows that $\{V(X_n) : n \geq 0\}$ in the case (iii) is bounded in probability. The case (iii) is a special case of S.P.E. with $q = 1$ so that the uniform quantitative bound of $\left\|P_\gamma^n(x,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}}$ for $\gamma \in \mathcal{Y}$ exists.

Lemma 4. *Suppose that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is S.P.E.. If the stochastic process $\{V_l(X_n) : n \geq 0\}$ is bounded in probability for some $l \in \{1, \dots, q\}$, then Containment is satisfied.*

Proof: We use the notation in Theorem 4.

From S.P.E., for $\gamma \in \mathcal{Y}$, let $\rho_{x,x'}(dy) = \delta\nu_\gamma(dy)$ (so $\rho_{x,x'}(\mathcal{X}) = \delta$) and $\triangle := C \times C$. So, $\epsilon^+ = \epsilon^- = \delta$.

Note that the matrix $I - A_m^{(\beta)}$ is a lower triangle matrix. Denote $(I - A_m^{(\beta)})^{-1} := (b_{ij}^{(\beta)})_{i,j=0,\cdots,q}$.

By the definition of $B_l^{(\beta)}(x,n)$,

$$\begin{aligned} B_l^{(\beta)}(x,n) =& \frac{\epsilon^+ \sum_{k=0}^l b_{lk}^{(\beta)} \int \pi(dy) W_k^\beta(x,y)}{S(l,n+1-m)^\beta + \sum_{j\geq n+1-m} (1-\epsilon^+)^{j-(n-m)} (S(l,j+1)^\beta - S(l,j)^\beta)} \\ \leq& \frac{\epsilon^+}{S(l,n+1-m)^\beta} \sum_{k=0}^l b_{lk}^{(\beta)} \int \pi(dy) W_k^\beta(x,y). \end{aligned}$$

By some algebra, for $k = 1, \cdots, q$,

$$\int \pi(dy) W_k^\beta(x,y) \leq 1 + \left(m(V_0) \prod_{i=0}^{k-1} a_i\right)^{-\beta} \left[V_k^\beta(x) + \pi(V_k^\beta)\right] \tag{40}$$

because $\beta \in (0,1]$. In addition, $m(V_0) \geq c_0$ so the coefficient of the second term on the right hand side is finite.

By induction, we obtain that $b_{10}^{(\beta)} = \frac{A_m^{(\beta)}(1)}{(1-A_m^{(\beta)}(0))^2}$, and $b_{11}^{(\beta)} = \frac{1}{1-A_m^{(\beta)}(0)}$. It is easy to check that $0 < b_{11}^{(\beta)} \leq \frac{1}{\delta}$.

By some algebra,

$$\begin{aligned} A_m^{(\beta)}(1) \leq & m^\beta + \sup_{(x,x')\in C\times C} \int R_{x,x'}(x,dy)R_{x,x'}(x',dy')W_1^\beta(y,y') \\ \leq & m^\beta + \sup_{(x,x')\in C\times C} \left[1 + (a_0 m(V_0))^{-\beta}(P_\gamma^m V_1^\beta(x) + P_\gamma^m V_1^\beta(x'))\right] \\ \leq & m^\beta + 1 + 2(a_0 m(V_0))^{-\beta}(\sup_{x\in C} V_1(x) + mb_0) \end{aligned}$$

because $P_\gamma^m V_1^\beta(x) \leq P_\gamma^m V_1(x) \leq V_1(x) + mb_0$. Therefore, $b_{10}^{(\beta)}$ is bounded from the above by some value independent of γ.

Thus,

$$\begin{aligned} B_1^{(\beta)}(x,n) \leq & \frac{\delta}{S(1,n+1-m)^\beta}\left(b_{10}^{(\beta)}\int \pi(dy)W_0^\beta(x,y) + b_{11}^{(\beta)}\int \pi(dy)W_1^\beta(x,y)\right) \\ \leq & \frac{\delta}{(n+1-m)^\beta}\left(b_{10}^{(\beta)}\pi(C) + b_{11}^{(\beta)}\left[1 + (a_0 m(V_0))^{-\beta}(V_1^\beta(x) + \pi(V_1^\beta))\right]\right). \end{aligned}$$

Therefore, the boundedness of the process $\{V_1(X_k) : k \geq 0\}$ implies that the random sequence $B_1^{(\beta)}(X_n,n)$ converges to zero uniformly on $\mathcal{X}$ in probability. Containment holds. □

Let $\{Z_j : j \geq 0\}$ be an adaptive sequence of positive random variables. For each j, Z_j will denote a fixed Borel measurable function of X_j. τ_n will denote any stopping time starting from time n of the process $\{X_i : i \geq 0\}$ i.e. $[\tau_n = i] \subset \sigma(X_k : k = 1,\cdots,n+i)$ and $\mathbb{P}(\tau_n < \infty) = 1$.

Lemma 5 (Dynkin's Formula for adaptive MCMC)**.** *For $m > 0$, and $n > 0$,*

$$\mathbb{E}[Z_{\tilde{\tau}_{m,n}} \mid X_m, \Gamma_m] = Z_m + \mathbb{E}[\sum_{i=1}^{\tilde{\tau}_{m,n}}(\mathbb{E}[Z_{m+i} \mid \mathcal{F}_{m+i-1}] - Z_{m+i-1}) \mid X_m, \Gamma_m]$$

where $\tilde{\tau}_{m,n} := \min(n, \tau_m, \inf(k \geq 0 : Z_{m+k} \geq n))$.

Proof:

$$Z_{\tilde{\tau}_{m,n}} = Z_m + \sum_{i=1}^{\tilde{\tau}_{m,n}}(Z_{m+i} - Z_{m+i-1}) = Z_m + \sum_{i=1}^{n}\mathbb{I}(\tilde{\tau}_{m,n} \geq i)(Z_{m+i} - Z_{m+i-1})$$

Since $\tilde{\tau}_{m,n} \geq i$ is measurable to $\mathcal{F}_{m+i-1}$,

$$\begin{aligned} \mathbb{E}[Z_{\tilde{\tau}_{m,n}} \mid X_m, \Gamma_m] =& Z_m + \mathbb{E}[\sum_{i=1}^{n}\mathbb{E}[Z_{m+i} - Z_{m+i-1} \mid \mathcal{F}_{m+i-1}]\mathbb{I}(\tilde{\tau}_{m,n} \geq i) \mid X_m, \Gamma_m] \\ =& Z_m + \mathbb{E}[\sum_{i=1}^{\tilde{\tau}_{m,n}}(\mathbb{E}[Z_{m+i} \mid \mathcal{F}_{m+i-1}] - Z_{m+i-1}) \mid X_m, \Gamma_m]. \end{aligned}$$

□

Lemma 6 (Comparison Lemma for adaptive MCMC)**.** *Suppose that there exist two sequences of positive functions* $\{s_j, f_j : j \geq 0\}$ *on* $\mathcal{X}$ *such that*

$$\mathbb{E}[Z_{j+1} \mid \mathcal{F}_j] \leq Z_j - f_j(X_j) + s_j(X_j). \tag{41}$$

Then for a stopping time τ_n *starting from the time* n *of the adaptive MCMC* $\{X_i : i \geq 0\}$*,*

$$\mathbb{E}[\sum_{j=0}^{\tau_n - 1} f_{n+j}(X_{n+j}) \mid X_n, \Gamma_n] \leq Z_n(X_n) + \mathbb{E}[\sum_{j=0}^{\tau_n - 1} s_{n+j}(X_{n+j}) \mid X_n, \Gamma_n].$$

Proof: From Lemma 5 and Eq. (41), the result can be obtained. □

The following proposition shows the relations between the moments of the hitting time and the test function V-modulated moments for adaptive MCMC algorithms with S.P.E., which is derived from the result for Markov chain in [17, Theorem 3.2]. Define the *first return time* and the *ith return time* to the set C from the time n respectively:

$$\tau_{n,C} := \tau_{n,C}(1) := \min\{k \geq 1 : X_{n+k} \in C\} \tag{42}$$

and

$$\tau_{n,C}(i) := \min\{k > \tau_{n,C}(i-1) : X_{n+k} \in C\} \text{ for } n \geq 0 \text{ and } i > 1. \tag{43}$$

Proposition 8. *Consider an adaptive MCMC* $\{X_i : i \geq 0\}$ *with the adaptive parameter* $\{\Gamma_i : i \geq 0\}$*. If the family* $\{P_\gamma : \gamma \in \mathcal{Y}\}$ *is* S.P.E.*, then there exist some constants* $\{d_i : i = 0, \cdots, q-1\}$ *such that at the time* n*, for* $k = 1, \cdots, q$*,*

$$\begin{aligned}\frac{c_{q-k}\mathbb{E}[\tau_{n,C}^k \mid X_n, \Gamma_n]}{k} \leq & \mathbb{E}[\sum_{i=0}^{\tau_{n,C}-1} (i+1)^{k-1} V_{q-k}(X_{n+i}) \mid X_n, \Gamma_n] \\ \leq & d_{q-k}(V_q(X_n) + \sum_{i=1}^{k} b_{q-i}\mathbb{I}_C(X_n))\end{aligned}$$

where the test functions $\{V_i(\cdot) : i = 0, \cdots, q\}$*, the set* C*,* $\{c_i : i = 0, \cdots, q-1\}$*, and* $\{b_i : i = 0, \cdots, q-1\}$ *are defined in the* S.P.E.*.*

Proof:

$$\sum_{i=0}^{\tau_{n,C}-1} (i+1)^{k-1} \geq \int_0^{\tau_{n,C}} x^{k-1} dx = k^{-1}\tau_{n,C}^k.$$

Since $V_{q-k}(x) \geq c_{q-k}$ on $\mathcal{X}$,

$$\mathbb{E}[\sum_{i=0}^{\tau_{n,C}-1} (i+1)^{k-1} V_{q-k}(X_{n+i}) \mid X_n, \Gamma_n] \geq \frac{c_{q-k}}{k}\mathbb{E}[\tau_{n,C}^k \mid X_n, \Gamma_n]. \tag{44}$$

So, the first inequality holds.

Consider $k = 1$. By S.P.E. and Lemma 6,

$$\mathbb{E}[\sum_{i=0}^{\tau_{n,C}-1} V_{q-1}(X_{n+i}) \mid X_n, \Gamma_n] \leq V_q(X_n) + b_{q-1}\mathbb{I}_C(X_n). \tag{45}$$

So, the case $k=1$ of the second inequality of the result holds.

For $i \geq 0$, by S.P.E.,

$$\begin{aligned}
&\mathbb{E}[(i+1)^{k-1}V_{q-k+1}(X_{n+i+1}) \mid X_{n+i}, \Gamma_{n+i}] - i^{k-1}V_{q-k+1}(X_{n+i}) \\
\leq&(i+1)^{k-1}\left(V_{q-k+1}(X_{n+i}) - V_{q-k}(X_{n+i}) + b_{q-k}\mathbb{I}_C(X_{n+i})\right) - i^{k-1}V_{q-k+1}(X_{n+i}) \\
\leq&-(i+1)^{k-1}V_{q-k}(X_{n+i}) + \tilde{d}\left(i^{k-2}V_{q-k+1}(X_{n+i}) + (i+1)^{k-1}b_{q-k}\mathbb{I}_C(X_{n+i})\right)
\end{aligned}$$

for some positive $\tilde{d}$ independent of i.

By Lemma 6,

$$\begin{aligned}
&\mathbb{E}[\sum_{i=0}^{\tau_{n,C}-1}(i+1)^{k-1}V_{q-k}(X_{n+i}) \mid X_n, \Gamma_n] \leq \\
&\qquad\qquad \tilde{d}\mathbb{E}[\sum_{i=0}^{\tau_{n,C}-1} i^{(k-1)-1}V_{q-(k-1)}(X_{n+i}) \mid X_n, \Gamma_n] + b_{q-k}\mathbb{I}_C(X_n).
\end{aligned} \tag{46}$$

From the above equation, by induction, the second inequality of the result holds. □

Theorem 8. *Suppose that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is S.P.E. for $q > 2$. Then, Containment holds.*

Proof: For $k = 1, \ldots, q$, take large enough $M > 0$ such that $C \subset \{x : V_{q-k}(x) \leq M\}$,

$$\begin{aligned}
\mathbb{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M\right) =&\sum_{i=0}^{n}\mathbb{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{i,C} > n-i, X_i \in C\right) + \\
&\mathbb{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{0,C} > n, X_0 \notin C\right).
\end{aligned}$$

By Proposition 8, for $i = 0, \cdots, n$,

$$\begin{aligned}
&\mathbb{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{i,C} > n-i \mid X_i \in C\right) \\
\leq&\mathbb{P}_{(x_0,\gamma_0)}\left(\sum_{j=0}^{\tau_{i,C}-1}(j+1)^{k-1}V_{q-k}(X_{i+j}) > (n-i)^{k-1}M + \right. \\
&\qquad\qquad \left. c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1}, \tau_{i,C} > n-i \mid X_i \in C\right) \\
\leq&\mathbb{P}_{(x_0,\gamma_0)}\left(\sum_{j=0}^{\tau_{i,C}-1}(j+1)^{k-1}V_{q-k}(X_{i+j}) > (n-i)^{k-1}M + \right. \\
&\qquad\qquad \left. c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1} \mid X_i \in C\right) \\
\leq&\frac{\sup_{x\in C}\mathbb{E}_{(x_0,\gamma_0)}\left[\mathbb{E}_{(x_0,\gamma_0)}\left[\sum_{j=0}^{\tau_{i,C}-1}(j+1)^{k-1}V_{q-k}(X_{i+j}) \mid X_i, \Gamma_i\right] \mid X_i = x\right]}{(n-i)^{k-1}M + c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1}} \\
\leq&\frac{d_{q-k}\left(\sup_{x\in C}V_q(x) + \sum_{j=1}^{k}b_{q-j}\mathbb{I}_C(x)\right)}{(n-i)^{k-1}M + c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1}},
\end{aligned}$$

and

$$\mathbb{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{0,C} > n \mid X_0 \notin C\right) \leq \frac{d_{q-k}\left(V_q(x_0) + \sum_{j=1}^{k} b_{q-j}\mathbb{I}_C(x_0)\right)}{n^{k-1}M + c_{q-k}\sum_{j=0}^{n-1}(j+1)^{k-1}}.$$

By simple algebra,

$$(n-i)^{k-1}M + c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1} = \mathrm{O}\left((n-i)^{k-1}\left(M + c_{q-k}(n-i)\right)\right).$$

Therefore,

$$\begin{aligned}
&\mathbb{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M\right) \\
\leq\ & d_{q-k}\left(\sup_{x\in C\cup\{x_0\}} V_q(x) + \sum_{j=1}^{k} b_{q-j}\right) \\
&\left(\sum_{i=0}^{n}\frac{\mathbb{P}_{(x_0,\gamma_0)}(X_i \in C)}{(n-i)^{k-1}\left(M + c_{q-k}(n-i)\right)} + \frac{\delta_{C^c}(x_0)}{n^{k-1}\left(M + c_{q-k}n\right)}\right).
\end{aligned} \tag{47}$$

Whenever $q \geq 2$, k can be chosen as 2. While $k \geq 2$, the summation of L.H.S. of Eq. (47) is finite given M. But if $q = 2$ then just the process $\{V_0(X_n) : n \geq 0\}$ is bounded probability so that $q > 2$ is required for the result. Hence, taking large enough $M > 0$, the probability will be small enough. So, the sequence $\{V_{q-2}(X_n) : n \geq 0\}$ is bounded in probability. By Lemma 4, Containment holds. □

Remark 20. *In the proof, only* (A3) *is used.*

Remark 21. *If V_0 is a "nice" non-decreasing function of V_1, then the sequence $\{V_1(X_n) : n \geq 0\}$ is bounded in probability. In Theorem 10, we discuss this situation for certain simultaneously single polynomial drift condition.*

Theorem 9. *Suppose that $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is S.P.E. for $q = 2$. Suppose that there exists a strictly increasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that $V_1(x) \leq f(V_0(x))$ for all $x \in \mathcal{X}$. Then, Containment is implied.*

Proof: From Eq. (47), we have that $\{V_0(X_n) : n \geq 0\}$ is bounded in probability. Since $V_1(x) \leq f(V_0(x))$,

$$\mathbb{P}_{(x_0,\gamma_0)}\left(V_1(X_n) > f(M)\right) \leq \mathbb{P}_{(x_0,\gamma_0)}\left(f(V_0(X_n)) > f(M)\right) = \mathbb{P}_{(x_0,\gamma_0)}\left(V_0(X_n) > M\right),$$

because $f(\cdot)$ is strictly increasing. By the boundedness of $V_0(X_n)$, for any $\epsilon > 0$, there exists $N > 0$ and some $M > 0$ such that for $n > N$, $\mathbb{P}_{(x_0,\gamma_0)}\left(V_1(X_n) > f(M)\right) \leq \epsilon$. Therefore, $\{V_1(X_n) : n \geq 0\}$ is bounded in probability. By Lemma 4, Containment is satisfied. □

Consider the single polynomial drift condition, see [17]: $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + b\mathbb{I}_C(x)$ where $0 \leq \alpha < 1$. Because the moments of the hitting time to the set C is (see details in [17]), for any $1 \leq \xi \leq 1/(1-\alpha)$,

$$E_x\left[\sum_{i=0}^{\tau_C-1}(i+1)^{\xi-1}V(X_i)\right] < V(x) + b\mathbb{I}_C(x).$$

The polynomial rate function $r(n) = n^{\xi-1}$. If $\alpha = 0$, then $r(n)$ is a constant. Under this situation, it is difficult to utilize the technique in Theorem 8 to prove $\{V(X_n) : n \geq 0\}$ is bounded in probability. Thus, we assume $\alpha \in (0, 1)$.

Proposition 9. *Consider an adaptive MCMC $\{X_n : n \geq 0\}$ with an adaptive scheme $\{\Gamma_n : n \geq 0\}$. Suppose that (A1) holds, and there exist some positive constants $c > 0$, $> b > 0$, $\alpha \in (0, 1)$, and a measurable function $V(x) : \mathcal{X} \to \mathbb{R}+$ with $V(x) \geq 1$ and $\sup\limits_{x\in C} V(x) < \infty$ such that*

$$P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + b\mathbb{I}_C(x) \text{ for } \gamma \in \mathcal{Y}. \tag{48}$$

Then for $1 \leq \xi \leq 1/(1-\alpha)$,

$$\mathbb{E}_{(x_0,\gamma_0)}\left[\sum_{i=0}^{\tau_{n,C}-1}(i+1)^{\xi-1}V^{1-\xi(1-\alpha)}(X_{n+i}) \mid X_n, \Gamma_n\right] \leq c_\xi(C)(V(X_n)+1). \tag{49}$$

Proof: The proof applies the techniques in Lemma 3.5 and Theorem 3.6 of [17]. □

Theorem 10. *Suppose that (A2) and the conditions in Proposition 9 are satisfied, and there exists some constant $b' > b$ such that $cV^\alpha(x)\mathbb{I}_{C^c} > b'$ for all $x \in \mathcal{X}$. Then, Containment is implied.*

Proof: Using the same techniques in Theorem 8, we have that

$$\begin{aligned} & \mathbb{P}_{(x_0,\gamma_0)}\left(V^{1-\xi(1-\alpha)}(X_n) > M\right) \\ \leq \quad & c_\xi\left(\sup_{x\in C\cup\{x_0\}} V(x) + 1\right)\left(\sum_{i=0}^n \frac{P_{(x_0,\gamma_0)}(X_i\in C)}{(n-i)^{\xi-1}(M+n-i)} + \frac{\delta_{C^c}(x_0)}{n^{\xi-1}(M+n)}\right). \end{aligned} \tag{50}$$

Therefore, for $\xi \in [1, 1/(1-\alpha))$, the sequence $\left\{V^{1-\xi(1-\alpha)}(X_n) : n \geq 0\right\}$ is bounded in probability. Since $1 - \xi(1-\alpha) > 0$, the process $\{V(X_n) : n \geq 0\}$ is bounded in probability. By Lemma 4, Containment holds. □

5.7 Proof of Theorem 6

Before we prove Theorem 6, we recall [16, Lemma 4.2].

Lemma 7. *Let x and z be two distinct points in $\mathbb{R}^d$, and let $\xi = n(x - z)$. If $\langle \xi, m(y)\rangle \neq 0$ for all y on the line from x to z, then z does not belong to $\left\{y \in \mathbb{R}^d : \pi(y) = \pi(x)\right\}$.*

Consider the test function $V(x) = c\pi^{-s}(x)$ for some $c > 0$ and $s \in (0, 1)$ such that $V(x) \geq 1$. By some algebra,

$$\begin{aligned} P_\gamma V(x)/V(x) = & \int_{A(x)-x}\left(\frac{\pi^s(x)}{\pi^s(x+z)}\right)q_\gamma(z)\mu_d(dz) + \\ & \int_{R(x)-x}\left(1 - \frac{\pi(x+z)}{\pi(x)} + \frac{\pi^{1-s}(x+z)}{\pi^{1-s}(x)}\right)q_\gamma(z)\mu_d(dz), \end{aligned}$$

where the *acceptance region* $A(x) := \{y \in \mathcal{X} | \pi(y) \geq \pi(x)\}$, and the *potential rejection region* $R(x) := \{y \in \mathcal{X} | \pi(y) < \pi(x)\}$. From [21, Proposition 3], we have $P_\gamma V(x) \leq r(s)V(x)$ where $r(s) := 1 + s(1-s)^{-1+1/s}$.

Proposition 10. *Suppose that the target density π is exponentially tailed. Under Assumptions 1–4, Containment holds.*

Proof: Note that it is not difficult to check that for $s \in (0,1)$, $\pi(V) < \infty$ by utilizing Definition 2.

Consider $s \in [0, 1/2)$. Under Assumption 4, let

$$
\begin{aligned}
h(\alpha, s) = r'(s) + \frac{1}{(1-s)^2} - \\
\frac{\alpha}{1-s} \inf_{(u,\gamma) \in S^{d-1} \times \mathcal{Y}} \int_{C_{\delta,\Delta}(u,\epsilon)} |z| \left[e^{-\alpha s |z|} - e^{-\alpha(1-s)|z|} \right] q_\gamma(z) \mu_d(dz) \text{ and} \\
H(\alpha, s) = 1 + \int_0^s h(\alpha, t) dt
\end{aligned}
$$

where $\epsilon, \beta, \delta, \Delta$, and $C_{\delta,\Delta}(\cdot,\cdot)$ are defined in Assumption 4. So, $H(\beta\epsilon/3, 0) = 1$ and

$$
\frac{\partial H(\beta\epsilon/3, 0)}{\partial s} = h(\beta\epsilon/3, 0) \leq e^{-1} + 1 - \frac{\beta\epsilon(1-e^{-1})}{3} \inf_{(u,\gamma)\in S^{d-1}\times\mathcal{Y}} \int_{C_{\delta,\Delta}(u,\epsilon)} |z| \, q_\gamma(z)\mu_d(dz) < 0.
$$

Therefore, there exists $s_0 \in (0, 1/2)$ such that $H(\beta\epsilon/3, s_0) < 1$.

Denote $C(x) := x - C_{\delta,\Delta}(n(x), \epsilon)$ and $C^\top(x) := x + C_{\delta,\Delta}(n(x), \epsilon)$. For $|x| \geq 2\Delta$ and $y \in C(x) \cup C^\top(x)$, $|y| \geq |x| - \Delta \geq \Delta$ so $|n(y) - n(x)| < \epsilon/3$.

Since the target density $\pi(\cdot)$ is exponentially tailed and Assumption 2, for sufficiently large $|x| > K_1$ with some $K_1 > 2\Delta$, $\langle n(x), \nabla \log \pi(x) \rangle \leq -\beta$ and $\langle n(x), m(x) \rangle \leq -\epsilon$. Then there exists some $K_2 > K_1$ such that for $|x| \geq K_2$, $\langle n(y), m(y) \rangle \leq -\epsilon$ for $y \in C(x) \cup C^\top(x)$. Thus, $|\nabla \log \pi(y)| = \frac{\langle n(y), \nabla \log \pi(y) \rangle}{\langle n(y), m(y) \rangle} \geq \beta$. Moreover, $y = x \pm a\xi$ for some $\delta \leq a \leq \Delta$ and $\xi \in S^{d-1}$. So,

$$
\langle \xi, m(y) \rangle = \langle \xi - n(x), m(y) \rangle + \langle n(x) - n(y), m(y) \rangle + \langle n(y), m(y) \rangle < -\epsilon/3. \tag{51}
$$

Hence, by Lemma 7, for $|x| > K_2$,

$$
C(x) \cap \left\{ y \in \mathbb{R}^d : \pi(y) = \pi(x) \right\} = \emptyset \text{ and } C^\top(x) \cap \left\{ y \in \mathbb{R}^d : \pi(y) = \pi(x) \right\} = \emptyset.
$$

For $y = x + a\xi \in C^\top(x)$,

$$
\begin{aligned}
\pi(y) - \pi(x) &= \int_0^a \langle \xi, \nabla \pi(x + t\xi) \rangle \, dt \\
&= \int_0^a \langle \xi, n(\nabla \pi(x + t\xi)) \rangle \, |\nabla \pi(x + t\xi)| \, dt \\
&< -\frac{\epsilon}{3} \int_0^a |\nabla \pi(x + t\xi)| \, dt \leq 0
\end{aligned}
$$

so that $C^\top(x) \subset R(x)$. Similarly, $C(x) \subset A(x)$.

Consider the test function $V(x) = c\pi^{-s_0}(x)$ for some $c > 0$ such that $V(x) > 1$. By Assumption 1, for any compact set $C \subset \mathbb{R}^d$, $\sup_{x \in C} V(x) < \infty$.

For any sequence $\{x_n : n \geq 0\}$ with $|x_n| \to \infty$, there exists some $N > 0$ such that $n > N$,

$|x_n| > K_2$. We have

$$P_\gamma V(x_n)/V(x_n) = \int_{\{C(x_n)-x_n\}\cup\left\{C^\top(x_n)-x_n\right\}} I_{x_n,s_0}(z)q_\gamma(z)\mu_d(dz) + \int_{\{C(x_n)-x_n\}^c\cap\left\{C^\top(x_n)-x_n\right\}^c} I_{x_n,s_0}(z)q_\gamma(z)\mu_d(dz),$$

where

$$I_{x_n,s_0}(z) = \begin{cases} \frac{\pi^{s_0}(x_n)}{\pi^{s_0}(x_n+z)}, & z \in A(x_n)-x_n, \\ 1-\frac{\pi(x_n+z)}{\pi(x_n)}+\frac{\pi^{1-s_0}(x_n+z)}{\pi^{1-s_0}(x_n)}, & z \in R(x_n)-x_n. \end{cases}$$

For $z = a\xi \in C^\top(x_n) - x_n$ and $t \in (0, |z|)$, by Eq. (51)

$$\langle \xi, \nabla \log \pi(x_n + t\xi)\rangle = \langle \xi, m(x_n+t\xi)\rangle\, |\nabla \log \pi(x_n+t\xi)| < -\epsilon\beta/3.$$

So, by Assumption 4,

$$\frac{\pi(x_n+z)}{\pi(x_n)} = e^{\log \pi(x_n+z)-\log\pi(x_n)} = e^{\int_0^{|z|}\langle \xi, \nabla \log \pi(x_n+t\xi)\rangle dt} \le e^{-\beta\epsilon|z|/3} \le e^{-\beta\epsilon\delta/3} \le e^{-1}.$$

Similarly, for $z = -a\xi \in C(x_n) - x_n$,

$$\frac{\pi(x_n)}{\pi(x_n+z)} \le e^{-\beta\epsilon|z|/3} \le e^{-1}.$$

$t^{1-s_0} - t \le \frac{1}{1-s_0}t^{1-s_0} - t$. Since $t \to \frac{1}{1-s_0}t^{1-s_0} - t$ is an increasing function on $[0,1]$,

$$\begin{aligned} &\int_{\{C(x_n)-x_n\}\cup\left\{C^\top(x_n)-x_n\right\}} I_{x_n,s_0}(z)q_\gamma(z)\mu_d(dz) \\ \le &\int_{C(x_n)-x_n} \frac{1}{1-s_0}e^{-s_0\beta\epsilon|z|/3}q_\gamma(z)\mu_d(dz) + \\ &\int_{C^\top(x_n)-x_n}\left(1-e^{-\beta\epsilon|z|/3}+\frac{1}{1-s_0}e^{-(1-s_0)\beta\epsilon|z|/3}\right)q_\gamma(z)\mu_d(dz). \end{aligned}$$

On the other hand,

$$\begin{aligned} &\int_{\{C(x_n)-x_n\}^c\cap\left\{C^\top(x_n)-x_n\right\}^c} I_{x_n,s_0}(z)q_\gamma(z)\mu_d(dz) \\ \le & r(s_0)Q_\gamma\left(\{C(x_n)-x_n\}^c \cap \left\{C^\top(x_n)-x_n\right\}^c\right). \end{aligned}$$

Define $K_{x,\gamma}(t) := \int_{C(x)-x} e^{-t|z|}q_\gamma(z)\mu_d(dz) = \int_{C^\top(x)-x} e^{-t|z|}q_\gamma(z)\mu_d(dz)$, and

$$H_{x,\gamma}(\theta,t) := \frac{K_{x,\gamma}(t\theta)}{1-t} + K_{x,\gamma}(0) - K_{x,\gamma}(\theta) + \frac{K_{x,\gamma}((1-t)\theta)}{1-t} + r(t)(1-2K_{x,\gamma}(0)).$$

So,

$$P_\gamma V(x_n)/V(x_n) \le H_{x_n,\gamma}(\beta\epsilon/3, s_0).$$

Clearly, $K_{x,\gamma}(t) \leq 1/2$. For $0 \leq t < 1/2$,

$$\begin{aligned}
&\frac{\partial H_{x,\gamma}(\theta,t)}{\partial t} \\
=& r'(t)(1-2K_{x,\gamma}(0)) + \frac{K_{x,\gamma}(\theta t)+K_{x,\gamma}(\theta(1-t))}{(1-t)^2} + \frac{\theta}{1-t}\left(K^{'}_{x,\gamma}(\theta t) - K^{'}_{x,\gamma}(\theta(1-t))\right) \\
\leq& r'(t) + \frac{1}{(1-t)^2} - \frac{\theta}{1-t}\int_{C(x)-x}\left(e^{-\theta t|z|} - e^{-\theta(1-t)|z|}\right)|z|\, q_\gamma(z)\mu_d(dz) \\
\leq& h(\theta,t).
\end{aligned}$$

Since $H_{x,\gamma}(\theta,0)=1$, $H_{x,\gamma}(\theta,t) \leq H(\theta,t)$ for $0 \leq t < 1/2$. Thus, $H_{x_n,\gamma}(\beta\epsilon/3,s_0) \leq H(\beta\epsilon/3,s_0) < 1$ so $\limsup\limits_{|x|\to\infty}\sup\limits_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < 1$. By Corollary 2, Containment holds. □

Remark 22. *Jarner and Hansen [16] shows that if under Assumption 2 the target density is lighter-than-exponential tailed, then the random-walk-based Metropolis algorithms are geometrically ergodic. The technique in Proposition 10 can be also applied to MCMC. So, even if the target density is exponentially tailed under some moment condition similar as Eq. (22), any random-walk-based Metropolis algorithm is still geometrically ergodic. In fact, our symmetry assumption ($q(x,y)=q(x-y)=q(y-x)$) is a little weaker than the assumption ($q(x,y)=q(|x-y|)$) of [16].*

Proof of Theorem 6: For (ii), by Proposition 10, Containment holds. Then ergodicity is implied by Containment and Diminishing Adaptation.

For (i), From Assumption 3, for any $\epsilon \in (0,\eta_1)$ and any $u \in S^{d-1}$,

$$\int_{C_{\zeta/2,\zeta}(u,\epsilon)} |z|\, q_\gamma(z)\mu_d(dz) \geq \frac{\iota\zeta \mathrm{Vol}(C_{\zeta/2,\zeta}(u,\epsilon))}{2}$$

where ι is defined in Eq. (20), ζ is defined in Assumption 3, $C_{a,b}(\cdot,\cdot)$ is defined in Eq. (21). The right hand side of the above equation is positive and independent of γ and u. Since target density is lighter-than-exponentially tailed, $\eta_2 := -\limsup_{|x|\to\infty}\langle n(x), \nabla\log\pi(x)\rangle = +\infty$ such that there is some sufficiently large β such that Eq. (22) holds. So, Assumption 4 is satisfied.

For (iii), adopting the proof of [12, Theorem 5], we will show that the simultaneous drift condition Eq. (14) holds. Denote

$$R(g,x,y) := g(y) - g(x) - \langle \nabla g(x), y-x\rangle\,.$$

Consider the test function $V(x) := 1 + f^s(x)$ where $f(x) := -\log\pi(x)$ for $\frac{2}{m}-1 < s < \min(\frac{2}{m}, \frac{3}{m}-2)$ where m is defined in Definition 3.

So,

$$P_\gamma V(x) - V(x) = P_\gamma f^s(x) - f^s(x) = \sum_{j=0}^{4} I_j(x,\gamma),$$

where M is defined in Assumption 5 and

$$\begin{aligned}
I_0(x,\gamma) :=& -sf^{s-1}(x)\,|\nabla f(x)|^2 \int_{R(x)-x\cap\{|z|\le M\}} \langle m(x), n(z)\rangle^2\,|z|^2\,q_\gamma(z)\mu_d(dz),\\
I_1(x,\gamma) :=& \int_{\{|z|\le M\}} R(f^s,x,x+z)q_\gamma(z)\mu_d(dz),\\
I_2(x,\gamma) :=& \int_{R(x)-x\cap\{|z|\le M\}} R(f^s,x,x+z)\frac{R(\pi,x,x+z)}{\pi(x)}q_\gamma(z)\mu_d(dz)\\
I_3(x,\gamma) :=& \int_{R(x)-x\cap\{|z|\le M\}} R(f^s,x,x+z)\,\langle \nabla f(x), z\rangle\, q_\gamma(z)\mu_d(dz)\\
I_4(x,\gamma) :=& \int_{R(x)-x\cap\{|z|\le M\}} \frac{R(\pi,x,x+z)}{\pi(x)}\,\langle \nabla f^s(x), z\rangle\, q_\gamma(z)\mu_d(dz).
\end{aligned}$$

By [12, Lemma B.4] and Assumption 5,

$$|I_1(x,\gamma)| = \mathrm{O}(|x|^{ms-2}),\ |I_2(x,\gamma)| = \mathrm{O}(|x|^{m(s+2)-4}),$$
$$|I_3(x,\gamma)| = \mathrm{O}(|x|^{m(s+1)-3}),\ |I_4(x,\gamma)| = \mathrm{O}(|x|^{m(s+2)-3}).$$

Note that the $\mathrm{O}(\cdot)$s in the above equations are independent of γ. Since $\frac{2}{m}-1 < s < \min(\frac{2}{m}, \frac{3}{m}-2)$, $|I_1(x,\gamma)|$, $|I_2(x,\gamma)|$, $|I_3(x,\gamma)|$ and $|I_4(x,\gamma)|$ converge to zero as $|x|\to\infty$.

By Assumption 2, for $\epsilon\in(0,\eta_1)$ (η_1 is defined in Eq. (19)), $\langle n(x), m(x)\rangle < -\epsilon$ as $|x|$ is sufficiently large. By Assumption 3, for sufficiently large $|x|$, for any $z\in C_{0,\zeta}(n(x),\epsilon)$ (ζ is defined in Assumption 3, ι is defined in Eq. (20), and $C_{\cdot,\cdot}(\cdot,\cdot)$ is defined in Eq. (21)),

$$-1 \le \langle m(x), n(z)\rangle = \langle m(x), n(x)\rangle + \langle m(x), n(z)-n(x)\rangle \le -\epsilon + \epsilon/3.$$

Thus,

$$\begin{aligned}
I_0(x,\gamma) \le& -\frac{4\epsilon^2 \iota s f^{s-1}(x)\,|\nabla f(x)|^2}{9}\int_{C_{0,\zeta}(n(x),\epsilon)} |z|^2\,\mu_d(dz)\\
=& -c_1 f^{s-1}(x)\,|\nabla f(x)|^2 \le c_2 f^{s-(2-m)/m}(x),
\end{aligned}$$

for some $c_1>0$ (independent of x) where $C_{0,\zeta}(n(x),\epsilon) = C_{0,\zeta}(u,\epsilon)$ for any $u\in S^{d-1}$.

So, there exist some $K>0$ and some $c_3>0$ such that $V(x)>1.1$ and $P_\gamma V(x) - V(x) \le -c_3 V^\alpha(x)$ for $|x|>K$, some $\alpha\in(0,1)$. Let $\tilde V(x) := V(x)\mathbb{I}(|x|>K) + \mathbb{I}(|x|\le K)$. So,

$$P_\gamma \tilde V(x) - \tilde V(x) \le -c_3\tilde V^\alpha(x) + c_3\mathbb{I}(|x|\le K).$$

Hence, by Theorem 5, Containment holds. □

5.8 Proof of Proposition 5

Note that in the proof of Theorem 6, some test function $V(x) = c\pi^{-s}(x)$ for some $s\in(0,1)$ and some $c>0$ is found such that S.G.E. holds.

To check Diminishing Adaptation, it is sufficient to check that both $\|\Sigma_n - \Sigma_{n-1}\|_M$ and $\left|\overline{X}_n - \overline{X}_{n-1}\right|$ converge to zero in probability where $\|\cdot\|_M$ is matrix norm.

We compute by standard algebraic manipulation that

$$\begin{aligned}
&\Sigma_n - \Sigma_{n-1}\\
=&\frac{1}{n+1}X_nX_n^\top - \frac{1}{n-1}\left(\frac{1}{n}\sum_{i=0}^{n-1} X_iX_i^\top\right) + \frac{2n}{n^2-1}\overline{X}_{n-1}\overline{X}_{n-1}^\top - \frac{1}{n+1}\left(X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right).
\end{aligned}$$

Hence,

$$\begin{aligned} & \|\Sigma_n - \Sigma_{n-1}\|_M \\ \leq \; & \tfrac{1}{n+1}\left\|X_n X_n^\top\right\|_M + \tfrac{1}{n-1}\left\|\tfrac{1}{n}\textstyle\sum_{i=0}^{n-1} X_i X_i^\top\right\|_M + \tfrac{2}{n}\left\|\overline{X}_{n-1}\overline{X}_{n-1}^\top\right\|_M + \\ & \tfrac{1}{n+1}\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M . \end{aligned} \tag{52}$$

indent To prove $\Sigma_n - \Sigma_{n-1}$ converges to zero in probability, it is sufficient to check that $\left\|X_n X_n^\top\right\|_M$, $\left\|\frac{1}{n}\sum_{i=0}^{n-1} X_i X_i^\top\right\|_M$, $\left\|\overline{X}_{n-1}\overline{X}_{n-1}^\top\right\|_M$ and $\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M$ are bounded in probability.

Since $\limsup\limits_{|x|\to\infty}\langle n(x), \nabla\log\pi(x)\rangle < 0$, there exist some $K > 0$ and some $\beta > 0$ such that

$$\sup_{|x|\geq K}\langle n(x), \nabla\log\pi(x)\rangle \leq -\beta.$$

For $|x| \geq K$, $\frac{\log\pi(y)-\log\pi(x)}{(r-1)|x|} \leq -\beta$ where $r > 1$ and $y = rx$, i.e. $\left(\frac{\pi(y)}{\pi(x)}\right)^{-s} \geq e^{s\beta\frac{r-1}{r}|y|}$. Taking $x_0 \in \mathbb{R}^d$ with $|x_0| = K$, $V(x) = c\pi^{-s}(x_0)\left(\frac{\pi(x)}{\pi(x_0)}\right)^{-s} \geq cae^{s\beta\frac{r-1}{r}|x|}$ for $x = rx_0$, $r > 1$, and $a := \inf\limits_{|y|\leq K}\pi^{-s}(y) > 0$, because of Assumption 1. If $r \geq 2$ then $\frac{r-1}{r} \geq 0.5$. Therefore, as $|x|$ is extremely large, $V(x) \geq |x|^2$. We know that $\sup_n \mathbb{E}[V(X_n)] < \infty$ (See Theorem 18 in [23]).

Since $\left\|X_n X_n^\top\right\|_M := \sup\limits_{|u|=1} u^\top X_n X_n^\top u \leq \sup\limits_{|u|=1}|u|^2|X_n|^2 \leq |X_n|^2$, $\left\|X_n X_n^\top\right\|_M$ is bounded in probability.

Obviously,

$$\left\|\frac{1}{n}\sum_{i=0}^{n-1} X_i X_i^\top\right\|_M \leq \frac{1}{n}\sum_{i=0}^{n-1}\left\|X_i X_i^\top\right\|_M.$$

Then, for $K > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=0}^{n-1}\left\|X_i X_i^\top\right\|_M > K\right) \leq \frac{1}{K}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\left\|X_i X_i^\top\right\|_M\right] \leq \frac{1}{K}\frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[|X_i|^2\right] \leq \frac{1}{K}\sup_n \mathbb{E}[V(X_n)].$$

Hence, $\left\|\frac{1}{n}\sum_{i=0}^{n-1} X_i X_i^\top\right\|_M$ is bounded in probability.

$\left|\overline{X}_n\right| \leq \frac{1}{n+1}\sum_{i=0}^{n}|X_i|$. So,

$$\mathbb{P}(\left|\overline{X}_n\right| > K) \leq \frac{1}{K}\frac{1}{n+1}\sum_{i=0}^{n}\mathbb{E}[|X_i|] \leq \frac{1}{K}\sup_n \mathbb{E}[V(X_n)].$$

$\left|\overline{X}_n\right|$ is bounded in probability. Hence, $\left\|\overline{X}_{n-1}\overline{X}_{n-1}^\top\right\|_M$ is bounded in probability.

Finally,

$$\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M \leq 2|X_n|\left|\overline{X}_{n-1}\right|.$$

Therefore, $\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M$ is bounded in probability. $\square$

5.9 Proof of Lemma 1

For $u \in S^{d-1}$,

$$\int_{C_{\delta,\Delta}(u,\epsilon)} |z|\, g(|z|)\mu_d(dz) = \int_\delta^\Delta g(t)t^d dt \int_{\{\xi\in S^{d-1}\,:\,|\xi-u|<\epsilon/3\}} \omega(d\xi).$$

where $\omega(\cdot)$ denotes the surface measure on S^{d-1}.

By the symmetry of $u \in S^{d-1}$, let $u = e_d := (\underbrace{0,\cdots,0}_{d-1},1)$. So, the projection from the piece $\{\xi \in S^{d-1} : |\xi - u| < \epsilon/3\}$ of the hypersphere S^{d-1} to the subspace $\mathbb{R}^{d-1}$ generated by the first $d-1$ coordinates is $d-1$ hyperball $\mathrm{B}^{d-1}(0,r)$ with the center 0 and the radius $r = \frac{\epsilon}{18}\sqrt{36-\epsilon^2}$. Define $f(z) = \sqrt{1-(z_1^2+\cdots+z_{d-1}^2)}$.

$$\begin{aligned}\omega\left(\left\{\xi \in S^{d-1} : |\xi-u|<\epsilon/3\right\}\right) &= \int_{\mathrm{B}^{d-1}(0,r)} \sqrt{1+|\nabla f|^2}dz_1\cdots dz_{d-1}\\ &= \frac{(d-1)\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}\int_0^r \frac{\rho^{d-2}}{\sqrt{1-\rho^2}}d\rho = \frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right).\end{aligned}$$

Hence,

$$\int_{C_{\delta,\Delta}(u,\epsilon)} |z|\, g(|z|)\mu_d(dz) = \frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right)\int_\delta^\Delta g(t)t^d dt. \tag{53}$$

Therefore, the result holds. □

5.10 Proof of Proposition 6

We compute that $\nabla\pi(x) = -\lambda n(x)\pi(x)$. So, $\langle n(x), \nabla\log\pi(x)\rangle = -\lambda$ and $\langle n(x), m(x)\rangle = -1$. So, the target density is exponentially tailed, and Assumptions 1 and 2 hold. Obviously, each proposal density is locally positive. Now, let us check Assumption 4 by using Lemma 1. Because

$$\mathrm{Vol}(\mathrm{B}^d(x,\Delta)) = \frac{\Delta^d\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2}+1)},$$

the function $g(t)$ defined in Lemma 1 is equal to $\frac{1}{\mathrm{Vol}(\mathrm{B}^d(x,\Delta))}$. η_1 defined in Eq. (18) and η_2 defined in Eq. (19) are respectively λ and 1. Now, fix any $\epsilon \in (0,1)$ and any $\delta \in (\frac{1}{\lambda},\infty)$. The left hand side of Eq. (26) is

$$\frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right)\int_\delta^\Delta g(t)t^d dt = \frac{d(d-1)}{2(d+1)\mathrm{Be}(\frac{d+1}{2},1/2)}\cdot\mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right)\cdot\Delta\left(1-\frac{\delta^{d+1}}{\Delta^{d+1}}\right),$$

where $\mathrm{Be}(x,y)$ and $\mathrm{Be}_r(x,y)$ are beta function and incomplete beta function, r is a function of ϵ defined in Lemma 1.

Once fixed ϵ and δ, the first two terms in the right hand side of the above equation is fixed. Then, as Δ goes to infinity, the whole equation tends to infinity. So, there exists a large enough $\Delta > 0$ such that Eq. (26) holds. By Lemma 1, Assumption 4 holds. Then, by Proposition 10, Containment holds. By Proposition 5, Diminishing Adaptation holds. By Theorem 1, the adaptive Metropolis algorithm is ergodic. □

Acknowledgements

We thank G. Fort and M. Vihola for helpful comments.

References

[1] C. Andrieu and Y.F. Atchadé. On the efficiency of adaptive MCMC algorithms. *Elec. Comm. Prob.*, 12:336–349, 2007.

[2] C Andrieu and E Moulines. On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

[3] C. Andrieu and C.P. Robert. Controlled MCMC for optimal sampling. . *Preprint*, 2001.

[4] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18: 343–373, 2008.

[5] Y.F. Atchadé. A cautionary tale on the efficiency of some adaptive Monte Carlo schemes. *Ann. Appl. Prob., to appear*, 2010.

[6] Y.F. Atchadé and G. Fort. Limit Theorems for some adaptive MCMC algorithms with subgeometric kernels. *Bernoulli*, 16:116–154, 2010.

[7] Y.F. Atchadé and J.S. Rosenthal. On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli*, 11(5):815–828, 2005.

[8] Y.F. Atchadé, G. Fort, E. Moulines, and P. Priouret. Adaptive Markov Chain Monte Carlo: Theory and Methods. *Preprint*, 2009.

[9] Y. Bai. Convergence of Adaptive Markov Chain Monte Carlo Methods. *PhD thesis, Department of Statistics, University of Toronto*, 2009.

[10] Y. Bai, R.V. Craiu, and A.F. Di Narzo. A mixture-based approach to regional adaptation for MCMC. *J. Comp. Graph. Stat., to appear*, 2010.

[11] R.V. Craiu, J.S. Rosenthal, and C. Yang. Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. *J. Amer. Stat. Assoc.*, 104(488):1454–1466, 2009.

[12] G. Fort and E. Moulines. V-Subgeometric ergodicity for a Hastings-Metropolis algorithm. *Statist. Prob. Lett.*, 49:401–410, 2000.

[13] G. Fort and E. Moulines. Polynomial ergodicity of Markov transition kernels. *Stoch. Process. Appl.*, 103:57–99, 2003.

[14] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7: 223–242, 2001.

[15] H. Haario, E. Saksman, and J. Tamminen. Componentwise adaptation for high dimensional MCMC. *Comput. Stat.*, 20:265–274, 2005.

[16] S.F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stoch. Process. Appl.*, 85:341–361, 2000.

[17] S.F. Jarner and G.O. Roberts. Polynomial convergence rates of Markov Chains. *Ann. Appl. Probab.*, 12(1):224–247, 2002.

[18] K.L. Mengersen and R.L. Tweedie. Rate of convergences of the Hasting and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 1996.

[19] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* London: Springer-Verlag, 1993.

[20] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

[21] G.O. Roberts and J.S. Rosenthal. Two convergence properties of hybrid samplers. *Ann. Appl. Prob.*, 8:397–407, 1998.

[22] G.O. Roberts and J.S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.*, 16:351–367, 2001.

[23] G.O. Roberts and J.S. Rosenthal. Coupling and Ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.*, 44:458–475, 2007.

[24] G.O. Roberts and J.S. Rosenthal. Examples of Adaptive MCMC. *J. Comp. Graph. Stat.*, 18 (2):349–367, 2009.

[25] G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110, 1996.

[26] G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.*, 7:110–120, 1997.

[27] G.O. Roberts, J.S. Rosenthal, and P.O. Schwartz. Convergence propperties of perturbed Markov chains. *J. Appl. Prob.*, 35:1–11, 1998.

[28] J.S. Rosenthal. Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *J. Amer. Stats. Assoc.*, 90:558–566, 1995.

[29] J.S. Rosenthal. AMCMC: An R interface for adaptive MCMC. *Comp. Stat. Data Anal.*, 51: 5467–5470, 2007.

[30] E. Saksman and M. Vihola. On the ergodicity of the adaptive Metropolis algorithms on unbounded domains. *Ann. Appl. Prob., to appear*, 2010.

[31] M. Vihola. Grapham: Graphical models with adaptive random walk Metropolis algorithms. *Comp. Stat. Data Anal.*, 54:49–54, 2010.

[32] C. Yang. On the weak law of large number for unbounded functionals for adaptive MCMC. *Preprint*, 2008.

[33] C. Yang. Recurrent and ergodic properties of adaptive MCMC. *Preprint*, 2008.